DOCUMENT RESUME

ED 166 195 95 TM 007 252

AUTHOR	Mosher, Frederic A.
TITLE	Final Report on the Instrumentation Study.
INSTITUTION	Huron Inst., Cambridge, Mass.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Jun 76
CONTRACT	OEC-0-74-0394
NOTE	128p.; Figures 1 and 2, page 46a, were removed from this document prior to its submission to the ERIC Document Reproduction Service ; Submitted December 1977
EDRS PRICE	MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS	*Achievement Tests; Adults; Aptitude Tests; Compensatory Education Programs; Criterion Referenced Tests; Educational Assessment; Elementary Education; Grade Equivalent Scores; Item Analysis; Norm Referenced Tests; *Program Evaluation; *Scores; *Standardized Tests; Test Construction; *Testing Problems; *Test Interpretation; Test Validity
IDENTIFIERS	Army Alpha Examination; *Metropolitan Achievement Tests; Project Follow Through

ABSTRACT

Focusing on the use of the Metropolitan Achievement Test (MAT) with Project Follow Through evaluations, this report represents an extended discussion of the appropriateness of standardized tests for evaluation and policy research. First, the development of the MAT and its roots in the Army Alpha intelligence test are described in detail. The next section discusses metric problems--that is, the types of scores standardized tests yield. The report recommends the use of the publisher's standard score, citing reasons why raw scores, percentiles, and grade equivalents are subject to misinterpretation. Psychometric problems--those concerned with test validity--are also discussed. It is argued that the use of standardized tests is not now justified because the tests may measure test wiseness, achievement motivation, general ability, and white middle class values rather than specific skills that schools are trying to teach. In the final section, recommendations are made for designing and validating new tests to fit particular goals: assessment and evaluation, identification of effective programs, diagnosis and prescriptions, and educational needs assessment. The appendix contains a history of the Army Alpha, notes on the norm referenced/criterion referenced distinction in testing, and results of the MAT analyses regarding the Philadelphia Follow Through Planned Variation evaluation. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED166195

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

The Huron Institute
123 Mt. Auburn Street
Cambridge, Massachusetts 02138
617 661-9285

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FINAL REPORT ON THE

INSTRUMENTATION STUDY

Frederic A. Mosher

June 1976

Carnegie Corporation
of New York
437 Madison Avenue
New York, New York 10022
212 371-3200

Prepared for:
The Office of Education
Follow Through Evaluation Branch
Contract No. OEC-0-74-0394

The opinions and conclusions in this report are the responsibility of the author and not of the sponsoring agencies.

(Submitted December 1977.)

TM007 252

INTRODUCTION

This is a final report on the work The Huron Institute has carried out with respect to problems of measurement and instrumentation as they affect the evaluation of the Follow Through Planned Variation Experiment. In our proposal we observed that a central purpose of school programs concerned with disadvantaged children, including Follow Through, has been to raise their levels of academic achievement. The instruments most often used to assess whether the goals of these programs have been attained have been one or another of the nationally available standardized achievement tests. We pointed out that serious questions have been raised about the appropriateness of the ways these tests have been used, and about the suitability of the tests themselves, for such evaluations. The concerns have focused on two suggestions: first, that the scores provided by the publishers for transforming raw scores into more interpretable form and the ways these scores are handled statistically in the usual analyses could produce results which are ambiguous or contradictory depending upon which scores and techniques were chosen; and second, that the way in which the tests are designed may make them relatively insensitive to real differences between school programs and thus not very helpful for use in research concerned with discovering whether programs have different effects. We proposed to explore these concerns with data available from the FTPV experiment.

In this report we present an extended discussion of the uses of measurement in evaluation and policy research, and of the problems involved in the use of standardized tests for these purposes. We spell out the nature of the metrics provided by the tests and show how the results of evaluations might be affected by those characteristics and by the ways they are handled. However, we also suggest that whether or not these concerns might cause real problems is probably an empirical question depending upon the particular test and the conditions of its use. We then briefly describe an analytic and empirical check of these problems as they affect the Metropolitan Achievement Test, the one used in the FTPV evaluation.

We then go on to a discussion of the question of whether standardized tests in fact measure what they seem to. This consideration of the psychometric assumptions and processes underlying such tests leads us to the conclusion that there are a number of very serious problems with such tests, of which the problem of sensitivity to program differences is just one. We conclude with a discussion of the kinds of steps which might be taken if new tests were to be developed which would fit more closely the various purposes for which tests are now used.

The points made in this report are not ones which will be a total surprise to psychometricians. The best of them (Chronbach, 1970) have made them in much more sophisticated and thorough ways. However, they are made in ways which usually do not get through

to researchers or policy makers who use the tests on a day to day basis, and we have tried here to couple the technical points with the kinds of judgments and prescriptions about what ought to be done which the psychometric fraternity seems to have been less willing to stress.

I. BACKGROUND

It seems to be a law of life that we don't normally look closely at what we are doing until we are somehow surprised--perhaps particularly not until we are unpleasantly surprised. In the case of standardized tests of academic ability or achievement for over four decades there had been few surprises. Rooted in the persistent psycho-philosophical tradition of concern with human differences and supported by a powerful statistical technology, the tests represented one of the few lasting examples of a useful product from "scientific" psychology. To be sure there were complaints all along from some whose expectations of equity or access were frustrated by the specific results of the tests, and from those whose sense of taste or of the complexity of life was violated by the content and form of test items and by their implication that most outcomes are uni-dimensional in character. But their complaints carried very little weight with those who were caught up in and responsible for the testing movement. The undeniable success of their predictions and the force of their psychometric assumptions left them secure in the conviction that there were no alternatives that could measure anywhere nearly so well, so fairly, or so economically.

In the 1950's and early 60's the nation rediscovered poverty and racial injustice. Because one stimulus for the rediscovery came from the courts' interventions on school integration and

their aftermath, and because of the seemingly obvious relationships between income and jobs and between access to jobs and success in school, a good bit of the attention to injustice focused on the performance of groups - poor people and Blacks - in school. Performance tended to be measured in the most credible and comparable available currency, i.e., not surprisingly, standardized tests. The results confirmed that the groups who otherwise did worse in the society also showed lower measured academic achievement. The obvious conclusions were jumped to, and they became the rationale for the first really major Federal support for education at the elementary and secondary level. While school people might have preferred support with fewer strings attached, their political allies seemed to be delighted to have found in injustice and poverty the leverage they had previously lacked to open the Federal Treasury to the schools. And some cynics would suggest that politicians and the powers-that-be were delighted to be able to spend relatively smaller amounts of money on the schools as the hoped-for path to more equal opportunity rather than to face the politically more difficult, and more expensive, task of dealing directly with inequality.

Whatever the mixture of motives, Federal intervention in education had come, but it arrived at a time when the government was beginning to offer its use of the newer techniques of rational

management - program planning and budgeting, for instance - as its latest defense against the perennial concern about levels of expenditure and waste. The government, in return for the ability and responsibility to spend the taxpayers' money, would identify real needs, set objectives for meeting them, select the best available among the proposed approaches to achieving those objectives, and continually check to see whether the objectives were being met or could be met better. The measures which were used to identify the need, particularly standardized tests of achievement in the basic cognitive skills of reading and mathematics, also became the obvious candidates for use in assessing the effectiveness of various proposed solutions and in evaluating the overall impact of the new programs.

The enthusiasm for needs assessment, program evaluation, and policy research was variously embodied in Federal legislation and bureaucratic action. Title IV of the 1964 Civil Rights Act mandated a study of the nations' schools to document the extent of inequality of opportunity and of results - the study which became the Coleman report. Titles I and III of Elementary and Secondary Education Act of 1965 required evaluation of the effectiveness of the projects supported under those titles, and these provisions were strengthened and extended in subsequent years. Later, in part in response to

some of the discouraging results of these earlier efforts, new, large-scale field "experiments" were established to try to test the effectiveness of some of the new programs in a more controlled way - such efforts as Experimental Schools, Head Start Planned Variation, and, of course, Follow Through Planned Variation.

Taken together these efforts represented an unprecendented attempt to pay attention to the actual outcomes of school programs and to try to sort out what does and does not work, what makes a difference, and what holds out the most hope for improving the performance of low-scoring children. The surprise, for many, came in the results. To simplify, and taking them overall, they seemed to say that once you control for the home background, the level of ability, or the entering performance of their students, existing differences in schools don't seem to be associated with differences in the measured outcomes of schooling. Furthermore, the experimental interventions and school innovations which have been tried don't seem to show lasting, consistent, and large impacts on students' achievement. The difficulty of finding "good" and "bad" schools, within the range of variation of schools as they are, severely shocked the conventional wisdom. The disappointing showing of the more innovative efforts of some of the best and the brightest people in education was an embarassment. The surprise was sufficient to

motivate reconsideration of a great many established notions and procedures.

We will not try here to review all of the avenues explored in the search for explanations or excuses. They have ranged from sorry attempts to place full blame on the recalcitrance of the children themselves to sophisticated demonstrations that the questions have not even really been tested yet. For purposes of this study it is sufficient to say that one of the centers of attention in the search for a culprit has been the output measures themselves. The attack has taken two forms.

The first says, in effect, "We weren't trying to do that." That is, the goals of the experimental programs had to do with something other than cognitive skills - usually with things like motivation, self-confidence, and self respect. These are seen as being prior requirements for academic work, or for successful performance in adult life, and it is suggested that it is not fair to evaluate programs trying to produce them with measures of narrow cognitive skills. This line is suggested most often with respect to programs aimed at pre-schoolers or the early grades.

The second confronts the standardized tests on their own cognitive turf and suggests either that they have been misused in these evaluations, or that they don't measure what they purport

to measure in an appropriate way, or both. In this report, when we wish to refer to more fundamental issues of the tests' validity, we will use the term "psychometric problems," because we hope to show that the psychometric assumptions underlying the tests may be faulty, or that the design steps based on them may involve dangerous shortcuts, so that the tests could be bad measures of the psychological constructs they are supposed to be tapping. Even if the psychometric assumptions and procedures are accepted as sound, the tests and measures which they generate have properties which can lead to misuse and misleading results unless they are recognized. Since most of these difficulties are concerned with the kinds of scores such tests yield, we have chosen to call them "metric problems."

We propose here to pursue only the second part of the attack on tests - the part which deals with their use as measures of cognitive skills - because we think that it has some merit and that we can contribute some clarification and make some progress on suggesting solutions for the problems. That is not to say that the other avenues are not also worth pursuing, only that we want to pursue on the streets we know.

Educational policy studies or evaluations are usually concerned with assessing the relative effectiveness of school programs. The standards of comparison vary. They include: comparison of the group being assessed with earlier performance of the same group or predictions based on it; typical performance as represented by national or other norms; the performance of a control group (or various statistical manipulations designed to substitute for the lack of a control group); or the performance of a group exposed to a competing treatment. Whatever form the comparison takes, it requires that students' performance be measured before and after, or at least after, exposure to the programs being assessed and that some summary measures

of the outcome be compared to the relevant standard. Since educational policy research is usually concerned with school programs, these measurements are often taken at the beginning and again at the end of the school year. The measures employed usually include one of the widely used tests of achievement in school subjects. As indicated, this report considers both the metric and the psychometric problems which we think raise doubts about the use of such tests for these purposes.

II. STANDARDIZED NORM-REFERENCED ACHIEVEMENT TESTS

Before we turn to an explanation of the problems involved in the use of these tests for a policy research we need to describe the ways such tests are developed and scored. The Metropolitan Achievement Test (MAT), published by Harcourt Brace Jovanovich, Inc., is typical of the better tests of this sort. It has been used in the FTPV National Evaluation since the spring of 1972, and for the most part we will refer to it here when specific examples are used.

In general the people who make nationally used achievement tests are trying to measure school skills - "the tests should reflect what is being taught in the classroom," and "should evaluate the extent to which the intended outcomes of instruction have been attained" (HBJ, June 1971, No. I, p. 3). A typical test provides a number

of batteries of sub-tests, each battery meant to be used at from one to three grade levels (see Table I for the six batteries of the MAT). The sub-tests which make up a battery include a number designed to assess various aspects of pre-reading and reading and mathematics, at the earlier levels, adding more specific subjects (spelling, science, social studies, language, etc.) for the later grades. The MAT starts in its Primer Level with Listening for Sounds, Reading (letters, simple words and sentences), and Numbers, and moves on the Primary I to Word Analysis (basically concerned with decoding - this drops out after Primary II), Word Knowledge, Reading (comprehension, interpretation, inference), Math Computation, Math Concepts, Math Problem Solving. Language and Spelling begin at the Elementary Level, and Science and Social Studies at the Intermediate Level.

Each sub-test is made up of a number of items (the MAT's range from 35 to 95, with most in the 35-60 item range), usually in a multiple choice format. The writing of the items and the choice of the ones to go in the final version of the test involve an elaborate process, and understanding it is central to evaluating these tests.

The process starts with an extensive review of the textbooks (in the subject to be tested) used in elementary schools across the country, along with consultation with curriculum experts,

Table I

Batteries of the MAT
(1970 Ed.) (Forms F,G,H)

Battery	Used at Grades
Primer	End of Kindergarten, Beginning-Middle First Grade (No Form G of Primer)
Primary I	End of First Grade, Beginning Second Grade
Primary II	End of Second Grade, Beginning Third Grade
Elementary	End of Third Grade, Beginning and End of Fourth Grade
Intermediate	Beginning and End of Fifth and Sixth Grades
Advanced	Beginning and End of Seventh and Eighth Grades, Beginning of Ninth Grade

reviews of various state and school system syllabi and objectives, and the recommendations of expert studies and commissions. The topics covered in these sources are tabulated along with indicators of the amounts of space and time devoted to them in each source, and these are, in a sense, averaged out and translated into a set of instructions to item writers (often classroom teachers or others close to the particular subjects) about what topics, content, and skills (often including process, or "intellectual abilities and skills," objectives of the sort defined by Bloom (1956), as well as more specific knowledge) should be covered, thereby setting a frame for the test, determining what relative emphases or weights (numbers of items) should be given the various objectives in the final version.

The objectives are chosen to be a fair representation of what schools in general are trying to do. Subject matter or nomenclature which are special to one text or curriculum are avoided. There is an attempt to keep up with shifting trends, however. The MAT manuals indicate that they anticipated that by the time their new version would be used very few schools would be using the "traditional" approaches to mathematics, and their specifications therefore called for items consistent with the objectives and nomenclature of "modern" math.

A large pool of items is written, and they are edited by experienced test editors, who try to make sure that the content of the items is consistent with the objectives they are supposed to measure (that they have "face validity"), that they are correct, that they aren't ambiguous unless they are supposed to be, and that they don't fall prey to any number of unwitting traps (like giving away the correct answer by internal clues which would not require knowledge of the specific material ostensibly tapped by the item, etc. (Wesman, 1971)). A great many more items are prepared than will be needed in the final test, and they are made up into trial tests, using trial instructions and trial timing, etc. These tests are then given to a fair sized trial sample of children in schools at the appropriate grades (the MAT used about 45,000 pupils in grades 1-8 from 22 school systems in 14 states (HBJ, June 1971, No. 6)). These samples are usually not random, but there is usually some attempt to make sure they represent the national diversity of American schools and children (though for some tests, not the MAT, there is an attempt to find "average" schools - in the sense of having mostly children of average "ability", as measured by standardized "ability" tests).

The results of this trial are used to select and revise the items which will go into the final version of the tests. Tables are

prepared for each item showing the proportion of the population who chose the correct answer and each of the "distractors" in the multiple choice. These proportions are also tabled for groups of children whose overall scores on the trial tests were quite high and those who were quite low (some test makers divide the populations into five or more groups according to their overall scores and show the proportions for each). A "good item" is considered to be one in which the high scoring group mainly chooses the correct alternative, and the successively lower groups tend to choose the distractors in larger and larger proportions. An item which low scorers tended to get correct, while high scorers missed it, would be an anomaly, and items which everyone got correct, or everyone failed, would not "discriminate" very well.

But we should say more about discrimination, and its companion, "difficulty," and their roles in item selection. The psychometric procedures used in these tests have their origin in the development of the "Army Alpha" and related tests which were devised around the time of World War I. The Alpha was developed to meet the need faced by the Army to find a quick way to assess the abilities of large and varied groups of recruits so that they could be channelled into military jobs at levels appropriate to their talent , time and resources would not be wasted on training people for jobs they couldn't do, and talented men would not be wasted on

doing jobs "beneath them." The test that was developed was considered to be a measure of intelligence or "native intellectual ability" (Yoakum and Yerkes, 1920, p.17), and it was hoped that it would measure ability to acquire new skills or to deal with new situations more than it assessed past achievement or instruction. The kinds of items chosen had their roots in the Binet-Simon intelligence test developed just after the turn of the century, and many of them were, or should have been, attempts to find tasks which could be assumed to be either equally familiar or equally unfamiliar to all of the test takers, so that differences in performance on them could be taken to be due to differences in some more stable underlying trait of individuals (the differences were widely considered to be genetically based, and the test makers were relatively innocent of or insensitive to the role class and cultural differences could play in exposure to the material contained in the items, though they did try, in developing the "non-verbal" "Beta" test to take into account the problems of the illiterate and the non-English speaking recruits). The purpose of the test was to discriminate among draftees and to make predictions about their suitability for, and future success in, various military jobs, so that differences in test scores could be used to sort and select among the men. Since the choices to be made were quite consequential for the men, and the Army - for instance the choice between being a foot soldier at risk of life or being, say, a desk officer - the discriminations had to be made as reliably as possible, i.e. a man shouldn't capriciously be identified

as officer material at one sitting of the test and tagged a doughboy at the next, and they should be valid, i.e. people who are predicted to do well should do well, and visa versa. The tests had to be fair, hence they were objective - they used items which had agreed-upon right and wrong answers and didn't rely on subjective judgements, as interviews or essay tests might have, and they were standardized - the conditions under which they were taken, the instructions, the time allotted, etc. - would be as alike as possible for everyone who took them.

While some of the rationale, and its mathematics, were refined well after the development of Army Alpha, the general argument about what produces a test which will be good at distinguishing among and ranking individuals is consistent with the notions developed then, and it runs like this:* The psychologists involved thought of psychological traits,skills, abilities, etc. as though they existed in some quantity - arrayed on a dimension or dimensions, in much the same way that physical attributes had quantity and could be dimensioned. In effect, each individual had a position on a dimension - his intelligence or level of skill - and so could each test item be considered to have a position on the same dimension, in the sense that some items were easy and hence low on the dimension, some difficult and high, etc. If a person were higher on the dimension than a particular item he

---

*See Appendix A for a more precise discussion of the history of the Alpha

should get it right; if lower he should fail it. If person and item were close on the dimension, and the notion of "error," some random process or processes such as motivation affecting the person's performance perhaps, were introduced, it might be that the person would get the item right on some occasions and not on others. If the person and the items were "on" the same point, and the error were random, on repeated trials (assuming no learning) the person would be above the item about as often as below, and would thus get it right about half the time.

If there were a lot of people clustered near a particular point on the dimension, it may be seen at least intuitively that the best way to split the group into those who are higher and those who are lower on the dimension would be to pick items whose position was in the middle of the group and make them into a test, with the number of items required to do the job reliably being a function of how closely clustered the group was on the dimension. These items could be identified as the ones which individuals in the group got right about half the time, if repeated, or which about half the group got right on one occasion. If a lot of items of this type are used there will be a slight tendency (its strength depending on the "distance" of the individual from the item) for those whose true position is just below the items to fail them more frequently than those who are just above. If the test score is

the number of items answered correctly, as the number of items increases one can have more and more confidence that the high scorers are truly higher on the dimension than the low scorers. For every item which is truly on the dimension there should be an association between getting that item correct and the chances of passing other items on the dimension or of getting a higher total score. That is, the items should "correlate" with each other and with the total score. In this assumed example the correlations would be low, because everyone is very close on the dimension and error has a real chance to operate, but still the people marginally above the items should have a marginally better chance of getting them correct and thus a better chance of having a higher total score. If items have crept into the pool which are not truly "on" the dimension, even though about half the group gets them correct, such items ought not to show the same correlation with the other items or with the total score (unless the dimension they are on is itself a correlated one, in which case they would be hard to distinguish statistically). If items are picked which are above or below the group (and hence more than half get them right if below, or wrong if above) the differential chances of the (true) higher and lower half of the group getting the item right will decrease (in proportion to the item's "distance" from the mid point), and it will take more and more items to identify

a difference reliably (going to the limit of everyone always getting the items right, or always wrong, in either case yielding no discrimination).

Now if a group of people is strung out along a dimension, and they all are given a group of items, it can be seen that the items might be ordered in their positions on the dimension by the proportion of the group who pass them. That is, easy items will lie below the group on the dimension and everyone will pass them. Nobody will pass very difficult ones, and the others will be passed in proportion to their positions on the dimension in relationship to the group, with due allowance for error. Furthermore, if there is only one dimension, once the items are ordered in <u>difficulty</u> (defined in the field somewhat perversely as the proportion of the group <u>passing</u> that item), it ought to be that if a person passes a harder item, he should pass all the easier ones. If not there must be error or perhaps the items are arrayed on more than one dimension. In the one dimensional case, another way of saying about the same thing as that passing a harder item implies that you will pass an easier item, is that the group's performances on the items will be correlated. In such a situation, if the goal is to identify how a group is spread along a dimension - to rank its members into as many ordered ranks as possible - it would be reasonable to pick items which have varying

difficulties - with quite a few items at each difficulty level. That in effect would be trying to place a cluster of items in the middle of each group of people in order to split each group in two, as in the group clustered around a point described above.

If the items aren't correlated very strongly with each other, and the test seems to be trying to measure some multidimensional characteristic of people in the group, then the group might be split up most effectively, at least in a first approximation, by picking items positioned on each dimension in the middle of the group, i.e. items with difficulty levels of about 50% (of course if you could identify the dimensions, say by factor analyzing the inter-item correlations, you could choose items of varying difficulty on each dimension and begin to build a number of uni-dimensional sub-tests).

Note that in deriving principles of item selection from this kind of model the reasoning has to turn around at some point. That is, it goes: If there were such underlying dimensions ordering people and items and if the pool of items we have written includes items which tap the dimensions at points appropriate for the group of people we are interested in, then we would expect to find at least some items which behave in the way you would predict from such assumptions, i.e. they might correlate with each other moderately and with total scores based on how many of the whole pool of items

were answered correctly, and they might vary in difficulty while being correlated with each other, or there might be a set of items which, while they didn't correlate very well with each other, still correlate pretty well with total scores and tend to have middle difficulty for the group. An actual pool of items could be searched for sets of items which behaved in these ways, and they could be made up into tests, and then they could be tried out on a comparable group to see if they still behaved in the same way (to make sure that the inter-item and item-total score correlations and the particular difficulty levels, didn't just happen by chance in the particular group selected for the try-out). Obviously the try-out group should be as representative of the groups for which the test is to be used as possible , or the test should be tried-out anew for each new group.

Now the Army Alpha test was being designed to predict performance in military jobs (and to exclude those who might be unfit for service), and in principle actual observations of performance on the jobs for an unselected but tested group could have been the criterion against which the items could have been selected and the test validated. That is, instead of just picking items which correlated with the total test score, the criterion for item selection could have been to take those items which discriminated between

(correlated with) those persons who did well and those who did poorly on the job. That was not the procedure used, however. For one thing, the people designing the tests were psychometricians who were striving for objectivity, and they were quite aware that job performance ratings were quite likely to be pretty subjective and unreliable. They were confident that the kinds of items they had written were reasonable examples of important skills, and they believed that the internal empirical evidence, given their assumptions, gave them more assurance that the tests they devised using these item selection rules were measuring something, and measuring it reliably, than they had that they or others could judge job performance. Their tests after all did replicate: they correlated pretty well with themselves when they were repeated on the same population; if the tests were split in half, randomly, the two halves correlated well with each other; and the tests correlated well with other tests made up in similar ways using different populations. These correlations were on the order of .75 to .9 or .95, clear indications that they were measuring the same, or closely related, things. To be sure, once the tests were made up they were also validated against military job ratings and other performance evaluations, and those relationships were high enough to justify the practical value of the tests (in general such relationships are

seldom over .5 and often rather lower - except where the prediction is to an academic-verbal rating, like grades in officer candidates school, say, or other written tests taken in the course of training), and even with the lower correlations the cutting points could be set at levels which would considerably reduce the chances that absolute incompetents would be taken into the service. How well they would have done on screening officer candidates, if they had had more time, might now be argued, but after some initial resistance the Army was reasonably pleased about the prospects.

The lower correlations with criterion performance were explained variously - the ratings themselves were unreliable, the jobs required skills or talents other than or in addition to those tapped by the tests - selection, reducing the "range of talent" of the group going into the criterion situation, also reduced the possibility of high correlations - but in any case the tests were essentially allowed to stand as true measures, and the issue became one of the degree of association between the attribute which the test was assumed to be measuring rather well, and the requirements of the job. (The Army over the years has also commissioned the development of a number of more specific, job related aptitude tests, which were designed on similar principles but which maintained a closer concern for using association with related performance as a criterion for item selection and test validation.)

The designers of the Army Alpha test, like other psychometricians, were interested in measurement - in being able to assign numbers to people or points on a psychological dimension in a way that would

mean something similar to the numbers found in physical measurement - and that interest led them to emphasize an additional assumption that was widespread at the time./ (Yerkes, 1921, pp.623ff.) It was: if various psychological attributes - like intelligence (which they considered to have a strong biological base) - of the whole of a population of human beings were measured on some true scale, the frequency distribution of people along points on the scale would take the shape of the normal ("Gaussian"-"bell-shaped") curve. It was a relatively simple step again to turn the reasoning around and to say: <u>if</u> the true distribution in the population is normal; if the raw score frequency distribution is obtained for the test for a particular sample of people (the numbers who get zero items right, one item right, etc.); if that distribution can be considered a representative sample of what it would be in the total population; then if the raw scores - 0,1,up to N (the number of items) are transformed by giving them the normal deviate (a measure of the distance in standard deviations from the mean of a standard normal curve) associated with the proportion of the area under the normal curve corresponding to the cumulative proportion of the population getting up to that score (i.e. 10% of the population might have gotten a score of 2 or less, 20% may have gotten 3 or less; the corresponding normal deviate for 2 would be about -1.28 and for 3 would be about -.84)

the resulting transformed numbers would be an "equal interval scale". That is, equal numerical differences between the transformed scores at different points of the scale would represent equal quantities (in the way that equal raw score differences might not - i.e. getting one more item right when you already have most of them may mean moving a much greater psychological distance than getting one more right when you are going from a raw score of 1 to 2, or visa versa). We should reemphasize, this logic depends on the assumption that the underlying population distribution is normal and that by transforming the scale numbers to, in effect, distribute the sample population normally you are bringing the scale "into line" with the true underlying scale. The normal deviates could be linearly transformed to any arbitrary mean (as 100 for I.Q.s) and standard deviation (15 or so for many I.Q. tests) without changing this scale property. The sample population was thus used to scale the items and the tests, not only in the sense of ordering the items according to difficulty, but also in terms of assigning new numbers to the levels of test performance, which numbers had a known and interpretable relationship to each other. It was quite a tour-de-force, and as we indicated, the power of the argument and the methods, reinforced by the seeming practical value of the results,

have had a lasting influence on test design.

To return to contemporary achievement tests, essentially the same principles of item selection as those derived for tests which were to be used to discriminate among individuals for selection and prediction are used for these tests of current academic attainment. To repeat, for maximum discrimination if items are highly correlated, one should select items of varying difficulty levels; if items are relatively uncorrelated, one should select items (which still should discriminate in the sense of correlating with the overall test score) whose difficulty level is around 50% (for multiple choice items, where guessing is possible, this percentage goes up into the 60's, the greater the number of alternatives the higher the percent) (Henryssen, 1971). The MAT designers sought items with high discrimination indices (in effect, high item-total score correlations), setting a minimum of about .20. That implies that the items used in the MAT are also at least modestly correlated with each other, and the designers had the intention of making reasonable discriminations all along the scale, so the range of item difficulties sought was also fairly broad (for the MAT they sought a range of .20 to .90) (HBJ, June 1971, No. 1, p. 14). The average item difficulties for the MAT were expected to be about .55 to .60 for the various sub-tests, and the distributions of item difficulties show

a considerably greater density around that middle range. This makes sense if the pupils are expected to bunch in that area and if the task of distinguishing among them is considered correspondingly more appropriate and difficult. The MAT designers strove for sub-test reliabilities (of the split-half sort) of around .90, a fairly typical requirement, probably even a bit on the stringent side.

The item data are examined carefully to select a set of items for making up the various sub-tests. A number of additional criteria come into play in the process. Most important is coverage of the content specified in the test frame. If the designers are short on items for particular content they will tinker with items having that content which have shown up badly on the difficulty or discrimination criteria and by changing distractors, etc. try to improve their performance. The MAT designers indicate that where there was conflict with the .20 minimum discrimination index guideline, content won out. They also caution users, however, that the items chosen to represent specific content users, however, that the items chosen to represent specific content <u>within</u> sub-tests are usually too few to represent kids' performance in those categories in a way that would be useful for diagnostic work, beyond an indication of something to look at more carefully and closely. Since the designers are interested in the growth of basic school skills over the school years they give preference to

items that show higher proportions of the tryout groups getting the items right in the higher grades among the grades taking the battery. Again, content takes precedence over this criterion, and the MAT designers report that this criterion was more likely to be satisfied in the lower grades, when children are receiving instruction and improving rapidly in these basic skills, than in the later elementary grades when the focus of instruction is on more specific content (at those grades they see, and sometimes accept, items whose proportion correct declines in the higher grade). At least some items are included in more than one test battery, to help test the continuity of measurement over the batteries.

The inspection of item data and editing and re-editing also involve a concern for various kinds of anomolies, and just how these are handled is in part buried in the lore of each testing organization. Obviously items which low scorers do well on relative to high scorers tend to be excluded. If it were noticed that there were items which seemed to be favoring particular school systems or curricular approaches they might be excluded, but it isn't clear that there is an explicit statistical search for such items, as opposed to an attempt to avoid them in the first place by writing fair items. There is increasing sensitivity to racial and cultural bias in the content of the test items and in increasing "representation"

in the item illustrations. There isn't much reported evidence that test makers exclude items which discriminate statistically among ethnic groups, but there is certainly room in this process, where one assumes there shouldn't be differences - as say between boys and girls in some areas, to exclude items which imply such differences, and to leave them in when the differences fit one's presuppositions.

Like the Army Alpha, the MAT and others like it do not use relationships with external criteria - like school grades or other performance indications of the skills to be measured - in the test design phase. Total scores on the test itself again play that role. Sometimes a known aptitude test is given at the same time as the new trial test as a basis for checking the representativeness of the sample or for use in weighting populations in the standardization phase, but the choice of items is essentially dependent upon internal statistical criteria and the face judgements about content made by experienced test developers and school people. Later, of course, the developed tests are correlated with other similar tests (usually showing strong relationships on the same order as the reliabilities) and with school grades (where positive but rather lower relationships are found - as in the Army Alpha, and they are explained in essentially the same way).

Following the item try-out and the choice of items to make up one or more forms of the sub-tests and the tests for each battery, the designers move to the standardization phase of their work. We have already used the words "objective" and "standardized" to apply to the content, format, and administration of the tests. "Standardized" has an additional, and more important, connotation for these tests. The designers want to identify a standard of performance against which to compare the performances of individual children or groups of children. They choose as their standard the performance of a representative population of comparable children--a reference group. Since their concern is with the outcomes of instruction and pupils' growth in achievement in school subjects they choose as their reference groups representative samples of school children at each grade level. They want to be able to say for each child how his or her performance in, say, arithmetic computation ranks in comparison with other children who have reached the same point in their schooling, or, turning the question around somewhat, to characterize his or her performance by saying which grade has an average/performance level which is (most) similar to the child's. If the designers were interested in something like general "mental maturity" they might choose age groups rather than grade level groups as their reference, but since they are concerned with school outcomes it seems reasonable for them to look at a group which

has had the "same" amount of schooling in terms of exposure to levels of curriculum rather than the same number of years of living. The degree of association between age and grade is obviously affected by school promotion policies. Currently in the U.S., with social promotion and tracking, the association may be closer than it used to be, say at the turn of the century, but the trend may be beginning to shift back to greater retention in grade.

The standardization phase involves the administration of the test batteries to a national sampling of schools and grades, the sample being designed to be as representative as possible. The MAT standardization sample involved over 100,000 children in 74 school systems , with at least 14,000 children at each of seven grades (2nd through 8th; the Primer was standardized and related to the Primary I Battery in a separate sampling of comparable size). Since children are assumed to be changing throughout the school year on the skills being measured, the sampling has to be done at roughly the same specified time in all the schools in order for the data to be comparable from school to school. Such tests are usually standardized in the fall of the school year, often in October, the second month. A number of the tests are now also standardized in the spring as well in order to give two points of comparison within each year. The MAT was standardized in April as well as in

October (there were over 50,000 children in each sample).

Most of the children involved in the standardization phase are given only the battery appropriate for their grade, but a smaller sample are either given a battery appropriate to an adjacent grade (higher or lower), or take both their own and an adjacent battery, or take some additional test which is given to students at adjacent grades as well. These steps are taken in order to provide data which can be used to put the scores from related sub-tests from the different batteries (i.e., all the math-computation sub-tests) onto one common scale. The MAT used a sample of 8,252 pupils in five school systems in their fall sample for this purpose, with children at each grade taking one or another of at least two, and in some cases three, of the MAT Batteries (they call this their "overlapping program"). The national standardization sample is checked to make sure its demographic characteristics match the national population as closely as possible, and usually all of the children are given some other standardized ability test to check on whether the sample is also representative of the population in terms of measured ability (the MAT standardization used the Otis-Lennon Mental Ability Test (OLMAT), and they report that the students at each grade level in the sample had a median I.Q. of 100). In the scaling

sub-sample, the ability tests are often used as a basis for selecting paired matched samples out of the groups who had taken the tests which were going to be equated or callibrated to each other.

If the test makers only wanted to give information about where a student stands among his or her grade peers (or those who would have been peers if the test were taken at the time of standardization) they could simply give the results of the standardization for each test at the grades in which it was taken. Students would be tested at the same time(s) of the year as the standardization(s) had been done and their scores could be interpreted by comparison to the parameters of the distribution of scores in the relevant standardization sample. That is, a score can be reported in terms of the proportion of the standardization sample whose scores were equal to or less than that score (its <u>percentile rank</u>) or in terms of its distance from the mean score of the standardizing sample, with the distance being measured in terms of the standard deviation of the sample. This transformed score, the difference between the raw score and the mean raw score, divided by the standard deviation of the raw score distribution, is called a <u>standard score</u> (or z-score). Such scores can be further transformed linearly to produce scores with a convenient but arbitrary mean, say 100, and an arbitrary standard deviation, say 10, or 15 as in I.Q.s. Or, as indicated

in the discussion of the Army Alpha, raw scores can be subjected to a "normalizing" transformation, and the scores can be reported as normal deviates, or some linear transformation of them (as, for instance, "T scores," whose mean is 50, with a standard deviation of 10). All of these alternatives can treat the parameters of the sample distribution as norms, in the sense of expected values which are to serve as standards for assessing performance, and this kind of reference to distributional standards is the source of the label norm referenced which is applied to these tests.

However, while supplying this kind of information is one of the aims of the test makers, they also want to say where a child's performance stands in terms of the expected development of a skill over the course of elementary school. In order to do this they need to relate the performances on comparable subtests from each of the batteries to some common scale whose range would represent the range of skill levels shown by children in kindergarten or the first grade through eighth or ninth grade or into high school. Most publishers provide for reporting results in terms of two somewhat different kinds of scales whose scores range across the grades in this way: the first are called, variously, scale scores or expanded standard scores; and the second are called grade equivalents (GE's). A basic assumption involved in the use of either of them is that the skill being measured - say, Word Knowledge, a MAT sub-test which is administered from first grade on - develops continuously in comparable

terms so that it is appropriate to think of it on one underlying scale with the populations of successive grades having successively higher mean performances on that scale, but also having distributions on the same scale around those means which would overlap with the distributions from other grades. The test actually given to eighth graders might be all but impossible for most first graders, but it is thought just to be measuring a higher level of the same skill (the analogy is sometimes drawn with thermometers used to measure temperatures in different settings - as in a cryogenics lab, a doctor's office, and a steel mill), and the overlap with intermediate tests can be used to relate the more distant tests to each other.

The scale scores are developed first. As we indicated, in the standardization phase the test makers find some way to see that comparable groups take the different forms and levels of the tests so that they can be callibrated to each other. For the MAT (H.B.J., Guidelines, No. 1, October 1972) adjacent levels of the test (Primary 1 and Primary II, Primary II and Elementary,...etc.) were administered to samples of children from pairs of grades (Primary I and II were given to 2nd and 3rd graders; II and Elementary to 3rd and 4th graders, etc.), where the alternative tests were assigned randomly to the various classes in the schools sampled, and where the resulting samples were further equated using OLMAT scores, so that

the populations which had taken the paired tests were arguably quite similar.

The development of the scales involves a number of steps. Since the Word Knowledge subtest, say, of the Primary II, is more difficult than that from the Primary I and has a different number of items, with different content, the first problem is to find a way of determining what level of performance on Primary II is equivalent to any particular performance on the Primary I. Since the groups who took each of the tests were supposed to be comparable, it is assumed that their distributions of skill on the two tests would be comparable and hence that if the cumulative percentages of the group achieving the various raw scores possible on each of the tests are plotted it will be possible to identify the raw scores from each test which are reached by the same proportion of the children who took them and to argue that these scores represent equivalent levels of test difficulty and equivalent performances (i.e. - If 30% of the children in the group who took the Primary I got 17 or fewer items correct, while 30% of those who took the Primary II got 13 or fewer correct, the raw scores 17 and 13 are taken to be equivalent to each other). Pairs of equivalent points identified in this way are plotted for the two tests, and a curve is fitted to the points, making it possible to read off equivalent raw scores on one test for any raw score on the other. The same process is carried out for the Primary II with the Elementary, the Elementary

with the Intermediate, etc., so that a table of equivalent raw scores is generated allowing a score on any one test to be related to a score on any other by moving through the equations with the tests in between.

A common scale for all the tests can be established using this table of equivalent raw scores by picking the raw score scale of any particular level and translating the scores of all the others on to that scale. However, the points of that scale, as with any raw score scale, will not necessarily have interval scale properties - i.e. equal numerical differences at different points on the scale will not necessarily represent equal differences in the underlying skill. The next step, to solve this problem, involves the use of the normalizing transformation discussed earlier. The test score distributions in the sample populations for each test level are obtained in terms of the equivalent raw score scale. Thus for each grade which took a comparable sub-test there will be a distribution of scores on the same scale, but the means and standard deviations will differ by grades (the means presumably increasing with increasing grade level). Each of these distributions is assumed to be normal in terms of some true underlying scale, though they won't necessarily be so on this equivalent raw score scale. It is possible to test whether any pair of these distributions could be normalized on the same underlying scale, albeit still with

different means and distributions (Thurstone, 1938). The test involves taking the percentile rank in each of the distributions for each score common to both of them and translating those ranks into normal deviates for each distribution. If the pairs of normal deviates are plotted and the points fall on a straight line, such a result would be consistent with the two distributions' being normal on a common scale. If the line is not straight, the distribution can't be normalized on the same scale. If the test is met, then it is reasonable to procede to the next step (in Thurstone's Absolute Scaling method, or Flanagan's (1939) related method). One of the grades is picked as a "base" (in the MAT, the earliest grade in which the particular sub-test appears), and its score distribution is normalized with a mean of zero and a standard deviation of one. By considering the points on the equivalent raw score scale / which the lower grade has in common with the next higher grade and the relationship between the normal deviate for those points in the two distributions, the "size" of the standard deviation of the second distribution can be determined with respect to that of the first distribution, and all of the scores in that distribution can be assigned tentative scores in the same units as the first distribution. The same procedure is used to put the second distribution and the next higher one on the same scale as the first two, and so on. The process yields two or more values for each score in the

raw score equivalent distribution, and these values are averaged and adjusted to remove reversals or other anomolies caused by imperfections in the curve fitting techniques used and irregularities in the data. The result is another common scale for all the tests providing numerical values in terms of the normal deviates of the base grade for each equivalent raw score, and through that, for each raw score value in the original tests. The raw scores and their equivalent normal deviates are plotted, and again smooth curves are fitted to the points, and new tables of raw score to normal deviate transformations for each test are generated from the curves.

These transformed scores now represent an arguably equal interval scale, but the numbers, since they are in the normal deviate units of the lowest of the successive distributions, from minus 2 or 3 to plus 6 or so, are not particularly convenient. They therefore undergo one last (this time linear) transformation to fit some arbitrary notion of useful numbers. The MAT scores are transformed to run from 0 to 150 over the grade range. This is done by "anchoring" the scale, at the low end, at the raw score representing the 2nd Percentile of the lowest grade distribution and assigning it the number 25, and, at the high end, in the 98th percentile score of the highest grade, giving it a score of 125. Using those two points, the remaining points could be assigned values by a

simple linear transformation.

Other test makers use comparable methods. The result is thought to be a scale of levels of development of each skill which has equal interval properties. The scores themselves do not give normative information, however, since, though their scale properties derive from reference to the original score distributions, those distributions have all been referred to one scale and the scores have lost the information they might have carried about their distance from the means of the distributions for particular grades. Because of this the publishers also supply information about the norms associated with the scale scores. To do this they go back to the distributions obtained in the total national standardization sample for each test at each grade. One way of interpreting a particular score is to ask which grade's "typical" performance it most resembles. That is the basis for the Grade Equivalent scale. It is generated by plotting the median scale scores for each grade in the standardization sample (In the MAT this was done for the October sample*) against the grade level (for these purposes the

*Actually, in the MAT (H.B.J. Guidelines No. 2, Feb. 1973) median raw scores were found for each grade level; these raw scores were converted, using the table of equivalent raw scores, into the raw scores of the sub-tests from two of the middle range battery levels (Elementary and Intermediate for full range sub-tests) and these converted raw scores were plotted against grade levels, and the curves were smoothed, yielding two plots for each sub-test. The Grade Equivalent raw scores generated in this way from the two plots seem to have been translated back into the raw score equivalent scale values and averaged, or otherwise

school year is divided into tenths, one tenth for each month of the nine in the school year and one for the summer - the "grade level" for October, the second month, is one tenth into the year - i.e. 1.1, 2.1, 3.1, etc.). A smooth curve is fitted to the points, and the grade level equivalent in tenths of a year can be read off the graph for any given scale score. It should be noted here that the use of curve fitting in this way involves the assumption that the growth in the skill being measured is regular between the times that are actually measured, and this suggests problems which we will return to later. Nevertheless, the Grade Equivalent scores provide another scale which runs across all of the school grades, and the scores provide more direct normative information, since it is quite easy to convert the score for a child at any given grade into the number of years or months s/he is "ahead" or "behind" his or her national peers (or rather, median peer).

The distributions by grade of the scale scores for the national standardization sample are also used to generate percentile

---

reconciled, and then the raw score equivalent table was used to provide grade level equivalent values for each raw score from each sub-test (not every grade level equivalent value has a corresponding raw score, however). The MAT tables its GE's in terms of standardized scores, so they must have worked back to standard scores through the raw to standard score conversion tables. It is not clear from their literature why the MAT publishers did not choose to work directly from their standard score scale, but it may have had something to do with fears about getting too many fitted curves away from the original data.

rank tables, showing the percentile rank which each scale or standard score has within the distribution of scores for pupils at a given grade level (at the times the tests were standardized - i.e., for the MAT, the tables are given for the beginning (October) and the end (April) of each year for which the particular battery is to be used - each scale score has a number of percentile ranks associated with it depending on the battery, sub-test, grade, and time of year for which it is tabled). The MAT also provides percentile ranks for scale scores obtained when the tests are given in the middle of the school year, and those are based on interpolation between the points established empirically in October and April. The percentile ranks give normative information about how much "above" or "below" his or her national peers a student is in terms of proportions of the reference group rather than how far ahead or behind in terms of years or months, as in the case of Grade Equivalent scores.

Most publishers also publish tables converting scale or standard scores to stanines. Stanines are scores derived from the normal distribution. They run from 1 to 9 (stanine is an elision of "standard nine"). They represent intervals whose width is 1/2 of a standard deviation, i.e. a score of 5 is assigned to scores which are within 1/4 of a standard deviation from the mean of the distribution on either side, and the scores of 4 and 6

are assigned to scores in the distribution which are between 1/4 and 3/4 standard deviations below or above the mean respectively, 3 and 7 between 3/4 and 1 1/14 below or above the mean respectively, and so on, with 1 and 9 including all scores more than 1 3/4 standard deviations below or above the mean. The stanine intervals contain differing percentages of the population distribution, corresponding to the areas under the normal curve contained by their boundaries. This fact makes them less directly interpretable for most users than are percentile scores, and since their original raison d'etre is no longer important (they were designed to fit the nine numerical possibilities in a column of an IBM punch card), they are not widely used. The MAT publishers do point out that the use of stanines does help "to avoid the overly precise connotations of two or three digit scores," and they can be assumed to represent equal intervals in the same way that the scale scores can.

Percentile Ranks and Grade Equivalents are tied to the distributions of test results in particular populations at a particular time, but they have the advantage of ease of interpretation in terms of the relationship of a score to expected performance. Scale scores don't offer this advantage, but in exchange their psychometric properties can be considered to be more stable. That is, they can be seen as representing an underlying scale of relative difficulty against which new populations can be normed, or along which the norms could change (i.e. this year's third graders might have a higher mean in scale score terms than did those in the standardizing population; scale scores would take on new percentile ranks with reference to this third grade, but the scale score values would not need to change; in that sense the scores are meant to be analogous to yardstick heights - the mean height of American adults is greater now than it was in 1918, but the scale of inches and feet has not changed its value). In practice, however, achievement tests don't seem to be re-normed against the scale scores in this way. Rather, when they are standardized and normed for a new edition, new scale scores are also published (partly because the items and tests are substantially revised at the same time; however, the new tests are usually equated to the old ones, in a smaller sample study, and the new norms could therefore be referred to the old scale, if one wanted to know whether the school age population

was improving, or declining, in skill).

III. METRIC PROBLEMS IN THE USE OF STANDARDIZED ACHIEVEMENT TESTS IN EVALUATION AND POLICY RESEARCH

With the forgoing background on the development and scoring of standardized achievement tests we can return to a consideration of the problems involved in using such tests in policy research and evaluation. This section deals with the problems we call "metric" because they are concerned with problems inherent in the kinds of measures yielded by the tests whether or not the tests are valid measures of the skills they purport to measure. That is, for the moment we will assume that the tests are valid and ask what pitfalls may nevertheless lie in using their scores. Many of the cautions we raise here have been raised in some form by the publishers themselves (see, for example, Wrightstone, Hogan, and Abbott), but they are widely ignored by researchers and policy makers. Where we can we will try to give some estimate of how serious the issues raised might be for practical results, and at the end of the section we will report the results of empirical tests of some of the questions. Then in the following section we will turn to problems which we think are even more fundamental - the problems of validity.

Perhaps the best way to see the kinds of trouble one can get into with these tests is to consider first how they are used in research and evaluation. The basic point is to use the tests to tell

what has happened to some group of students (or occasionally a student) who have been exposed to some treatment - i.e. a year in a particular school and classroom, a new reading curriculum, desegregation, and so on. As we indicated, to answer such a question, one must contrast what happens to the group, as measured by the tests either after or before and after the treatment, with what would have happened without the treatment, or if some other treatment had been administered. To do that one must have comparable measures on an appropriate non-treatment group of students. Sometimes this non-treatment group is an actual control group, comparable to the treatment group in as many respects other than the treatment as possible, and it is tested in the same way. Often, when the research is concerned with some general school program acting over periods of a year or more, the implicit contrast is simply with U.S. school children in general, as represented by the norms of a nationally standardized test. The two situations, comparison to controls or to norms, offer differing opportunities for error.

In the case of comparison to norms one is usually asking "Are the kids in this school, or program, doing as well as or better than kids in general at that grade?" or, sometimes, "Did they gain as much this year on the average as the average kid should gain?" Now, treated as descriptive questions, these are reasonably straightforward. It is appropriate to ask where these children

stand with respect to the distribution of scores of children at this grade, or where this school mean, say, stands in relationship either to the distribution of individual scores, or, if it were available, to the distribution of school means. Nevertheless, such a comparison to the norms does assume that the norms are accurate empirical descriptions of the distributions of scores in the school population at the grade and time in the school year for which the comparison is made. As we note below, under some circumstances this assumption is tenuous. In addition, however, if the point is to say something about how well the school or program is doing its job, then comparison to the norms used as a basis for answering the question involves another assumption, if the approach is to be "fair."

That assumption is that the population of children in the grade or program is typical - i.e. that it is a representative sample of the national population of school children in all respects other than those which have to do specifically with differences in schools or programs. The assumption is required because it is not usually assumed that all children are equally educable and that all of the variation in achievement is due to variation in school effectiveness. Thus, if schools should differ in the average ability and background of the children assigned to them it might not be reasonable to expect them all to perform at the national norm levels. In effect, however,

this/often imposed expectation is when school to school mean or median achievement scores are published for a district. Parents and the press tend to hold the schools accountable when school mean or median grade equaivalents fall below grade, or their percentiles fall below the 50th, and it is the fear of this "misunderstanding" and of the difficulty of correcting it that in part explains the lack of enthusiasm among school people for public reporting of such scores. Still, it might be argued that all children <u>are</u> equally educable, or at least that all are educable up to certain standards, and schools <u>might</u> be expected to bring their students (or the average of their students) up to grade level or to the 50th percentile (the irony of asking the schools to bring everyone above the median is only partial, since, while a stringent standard, it is <u>possible</u> to do so for a new group of children if the reference is to the distribution of scores in another sample and not to the median of the group itself - i.e. it is impossible, as opposed to unreasonable, only if the test is renormed on the group itself). In such a case the norms would be treated as though they were the <u>criteria</u> for school effectiveness or individual performance. As we will argue later in Section V, and to some extent in IV, we feel that it may be appropriate to try to set specific, or "absolute," criteria for school effectiveness, but it would be better to use tests which are explicity designed to justify such use (i.e. "criterion-referenced" tests, of a sort) rather than to try to recruit the norms on these

tests into such service.

The issue of the representativeness of the school or program populations, when the tests are to be used to assess school or program effectiveness, is largely a research design rather than a metric issue, though the attempts made to solve it do bring with them serious metric questions, as we shall suggest shortly. But first there is the question raised even by the straightforward comparison of observed scores to the norms. As we noted, achievement tests are standardized at one point, or occasionally at two points, in the school year. The tables of scale or raw scores which represent the expected median scores at each grade for each month during the school year (Grade Equivalents), or which have various percentile values at points in the year other than the times at which the tests were standardized, are generated by fitting smooth curves to the empirically obtained points and interpolating for the periods in between. Such curve fitting obviously assumes that the growth of the skill being measured is regular during the period for which the interpolation is made. But that assumption is not likely to be true for all skills. For instance, consider a test which is standardized at one point in the year, in the fall. It might be reasonable to expect that school skills on the average would be growing more rapidly during the school year than during the summer. Thus in fact they might reach a peak of accelleration in the spring

Figures 1 and 2 have been removed because of copyright restrictions.

and then level off relatively over the summer. In such an instance the line drawn between the fall points from one year to the next would fall below the spring points which would be found if an empirical sample were actually taken then (see Figures 1 and 2). Thus the tabled G.E. raw or scale score for a point in the spring would be lower than that which would be obtained empirically, and a student, or a group, who scored at what would be the empirical median score for that time would be given a higher G.E. by the table, i.e. he, she, or they would seem according to the table to have done better than they actually did. It is possible that other kinds of irregularities in the growth of skills could produce errors in the opposite direction, leading to underestimates of actual performance, or they could produce even greater overestimates. Percentile scores are usually tabled for fewer points in a year, but where they are based on interpolation the same kinds of errors could occur.

The magnitude of these possible errors can't be estimated without doing empirical norming at additional times and comparing the results to the interpolated values, and this has not been done, or at least it hasn't been reported in a clear way, though the publishers have been sufficiently concerned about the problem to warn of the possibility and to increase the number of norming points in their standardization (HBJ, May 1972, No. 21; HBJ, June 1971, No. 8; Wrightstone, Hogan, and Abbott, N.D.). The problem can be

avoided in evaluation or policy research by using tests which are normed at more than one time in the year and by testing only at the times in the year for which empirical norms are available (for most school studies the obviously most relevant norming and testing points would be as early in the fall and as late in the spring as possible).

As we noted, the MAT 1970 Edition was in fact normed at two times in the year - October and April. The beginning and end of year percentile scores (and stanines) are empirically based on those two points. The Grade Equivalent scores tabled for the MAT are, however, based on interpolations between the fall points only (HBJ, June 1971, No.8). If one looks at the Grade Equivalent values for the empirical median scores obtained in the spring for each sub-test in the MAT batteries for grades 1 through 8 (Op.Cit., pp. 9-11), of the 76 sub-tests, 42, or more than half, show GE's two months or more ahead of or behind (actually 21 two or more ahead and 21 two or more behind) the expected value of seven months into the year. Only 13 of the 76 are at exactly seven months, and three are half a year or more off (all of the latter behind the expected value and all in the seventh or eighth grade). The dispersion of the obtained values from the expected value increases from the first to the eighth grade, and, while there are values ahead of and behind those expected for each grade other than the first, where they are all ahead, the

values tend more to be ahead of the expected in the earlier grades and behind it in the later. Thus children measured in the spring will tend to seem (spuriously) to do well in the earlier years and badly in the later ones, and while the effect is not large relative to the standard deviation of the individual score distribution, it could be more of a problem with respect to the distributions of class or school means (where the size of standard deviation would be less, but the size of the spurious effect would remain the same).

For the MAT at both norming times, the mean value of the test scores in GE's is ahead of the median value in the early grades and falls behind in the upper grades, often by several months. The publishers point out that this reflects the fact that the distribution of GE's is, characteristically for this scale in all tests, often quite skewed, and it is also an artifact of their putting upper and lower limits on the GE's which can be given for the different batteries, which in itself would depress the means for the upper grades and inflate them for the lower. These facts suggest that the use of mean GE's for comparison of school or program results to national norms is particularly shaky.

To recap, where direct comparison of school or program outcomes to norms is justified by design considerations, metric concerns suggest that the comparisons be made only at points in the year at which the test used has been empirically normed and only in terms of the

scores for which empirically based norms are reported (i.e. it is risky to use GE's for MAT spring scores, while percentiles would have somewhat greater justification).

A number of techniques have been tried to deal with situations in which it is not appropriate to make a simple comparison of school or program results to the national norms, i.e. where one wants to make inferences about the effectiveness of the "treatment" and where it is clear that the student population does not represent a sample from the same population as the norming sample. These techniques all in one way or another involve, in effect, correcting the norm or the value expected for the group by taking into account the ways in which the group differs from the norming sample. In order to do that one must have additional knowledge about the association between the characteristics on which students or groups of students might differ and performance on the tests. That is, one would want to regress the achievement scores on all of the characteristics one could think of and measure which might have something to do with achievement. Then one could use all of the non-school characteristics in a multiple regression procedure to predict for any given school or program what its students should do on the tests given the values of their particular set of non-school characteristics and assuming the school or program characteristics were the same as the average. The difference between that value

predicted for a school or program and its actual value might then be an index of its relative effectiveness - i.e. given that all the relevant non-school factors had been controlled for in this way, the only reason for a school or program to differ from its predicted value would be something to do with the school or program itself, or chance, or error. Attributing the effect to particular aspects of the school or program would involve additional design and measurement complications, but one could argue at least that something to do with them was making a difference.

This, of course, is analogous to the procedure that was used in the Coleman study, and in other cross-sectional surveys, and it has often been suggested as a basic design for school accountability systems. In these instances there is often no direct reference to national norms since the regressions are performed on the data gathered within the study or the assessment scheme itself. The results are nevertheless often reported in terms of the metrics of the national norms and take some of their significance from them; and, while we wouldn't necessarily advocate it, it is not hard to see how comparable data could be provided, and studies done, using the tests' own national standardization samples.

There is a refinement on this design, which ought to be discussed below in the context of questions of change or growth in achievement rather than in the context of questions of relative achievement out-

comes, but which we choose to discuss here because it involves the same metric issue we want to raise about the use of regression techniques in comparing outcomes. (Besides, consideration of outcomes with concern for attributions of effectiveness to schools or programs obviously in any case involves an implicit concern with change - the differential change brought about by the differences in effectiveness which one wants to infer from the differing outcomes). It is the use of the regression of post-treatment scores on pre-treatment scores as a basis for controlling for differences in the composition of groups to be compared to the norms (or to each other). Differences between the obtained scores and those predicted on the basis of the regression may be seen as reflecting differences due to the intervening treatments, if one assumes that all of the effects of non-treatment differences are already "contained" in the pre-test scores and that they don't continue to exercise a differential effect during the "treatment" (as of course they could!).

We are oversimplifying the design considerations here, and perhaps lumping too many things together, but we do it because the metric problem we want to emphasize is simple and fundamental and applies to all variants of these procedures. It is that the techniques used all assume that the characteristics being measured and related are measured on an interval scale (Stevens, 1951), and since we have already shown that the various scores derived for achievement tests -

raw scores, scale scores, percentiles, and GE's - all represent non-linear transformations of each other, it is clear that, in principle, if one of them meets the requirements for being considered an interval scale the others cannot. The only scores for which there is an explicit case made that they should represent an interval scale are the scale or publishers' standard scores, and, as long as the psychometric assumptions underlying these tests are being accepted, we can see no good reason to employ scores other than the scale scores when regression techniques are to be used. The relationships among the score transformations are to be sure monotonic. The non-linearity may not be extreme, particularly in the middle ranges of the tests, and the degree of problem one might have can be tested for if you want to go back and look carefully at the publishers' standardization data. There may seem to be an advantage to working with GE's or percentiles if you can, since the differences you find will be in metrics which seem to have more directly interpretable significance. But we think the risk outweighs what anyway may be a specious advantage. Raw scores have the appeal of concreteness, but they can't be compared across different batteries, and it is all too clear, given the wide range of item difficulties, that equal raw score differences don't necessarily have equal meaning. That's why the test makers devised scale scores in the first place. If you use scale scores you can always try to

transform the results back into the units of the other scores, and we think the exercise would prove instructive (hint: you can't enter the tables with a scale score difference).

Turning now directly to the question of whether the students in a school or program have gained as much in a year as they should, we see that the issue is still largely one of the degree to which the students can be considered to resemble the norming sample. If they do they should be expected to gain on the median a year per year in Grade Equivalent terms (subject to the caveat noted above about the effects of interpolation), to hold at the 50th percentile for their median score, and to grow the same average number of scale score points per year as the national norming group. However, when one gets away from samples from the same population as the norming sample or, for individuals, away from the center of the distribution, the expectations, and the relationships among them, change. The changes are, again, essentially a matter of the non-linearity of the relationships among the alternative metrics, plus the inhomogeneity of the variances of some of the metrics over the school grades.

The classical example of this problem was the finding in the Coleman study that the average performance of students from low income homes and Blacks was a bit behind the national norms in grade equivalent terms at the first grade and that it seemed to fall further and further behind for each succeeding grade. Later, however, Coleman (Coleman and Karweit, 1970) observed that this finding was

essentially an artifact of the distribution of GE's and the way they are constructed, and that if normal growth were defined as maintaining one's position in the distribution relative to one's peers, poor kids and Blacks remained on the average about the same distance below the national norm (i.e. they kept the same average percentile score) from year to year. Coleman stressed the point that the distance behind in GE terms could change while the distance below in percentile terms could remain constant. In fact it had to change, and increase, if the percentile were to remain constant.

The reason for this is simply that the variance of GE's increases from grade to grade. This is partly because the range is necessarily restricted at the lower end, but that is not the whole reason, since the test makers also restrict it at the higher end. It seems mainly to reflect the increasing variability of achievement as measured by these tests, since the variance of standard scores also increases, as does that of raw scores from grade to grade within a battery. On the MAT the variance of Total Reading GE scores goes from .77 years at the first grade to 2.17 years at the eighth grade (declining to 2.00 years at ninth, perhaps because of the upper end restriction), thus more than doubling. This means that an eighth grader who tested at two years behind grade would, roughly, have the same percentile score as a first grader who was seven months behind.

In general in these kinds of achievement tests a "year's growth per year" is the empirical expectation only for children at the median. Children above the median have to grow more than a year in GE terms, and to increase more than the increase in the mean/score for the norming population, if they are to remain at the same percentile. Children below the median have to increase less on either scale to stay even in percentiles. Wrightstone, Hogan, and Abbott (N.D.) in discussing the alternative metrics which might be used for defining normal growth suggest that since scale scores can be argued to have equal interval properties it might be appropriate to consider change from year to year in the mean scale score of the norming population to be a reasonable standard for "normal" growth, since a child no matter where he started who moved that amount in scale scores could be said to be moving the "same" amount as the norming group. However they recognize that while this could be asserted as a standard, it would not represent an empirical expectation for the growth of a particular child or group of children. The relationship between scale scores and percentiles, while not quite as extreme, is similar to that between GE's and percentiles.

We think that while the debate about how and in what metric normal growth should be defined is interesting, and it has been useful in forcing somewhat greater attention to the nature of the metrics themselves, and while we share the bias of almost everyone

who understands them against GE's, and perhaps in favor of percentiles since they seem to offer a clear interpretation, it is basically misleading for the purposes of policy research and evaluation. We say that because what are required for asking whether a school or program has produced more or less change than was expected are empirical expectations about how kids who have particular starting points are likely to shift, on the average, over a year (or whatever period is of interest). The publishers' norms simply don't give such expectations, since they are cross-sectional. They don't look at the same kids at two or more different times.

There are problems of course in providing such change "norms," and not just because it would take a long time. In the MAT standardization a fairly representative sub-population of children was tested in the fall and then retested in the spring. We had a chance to look at some of the results of this exercise, which they called the "Fall-Spring Growth Study" (Beck, 1974). The most interesting thing for our purposes was a graph showing the growth in mean standard scores from fall to spring for each grade for three groups, identified on the basis of their fall scores - the top, middle, and bottom three stanines of the fall distribution. The lower group consistently showed the greatest gain in scale score terms, the middle group the next greatest, and the top group the least. Translated into GE's these differences tend to wash out or even reverse.

For the bottom group the spring scale score mean of one year is often higher than the fall mean for the lower group in the next grade. If these were the same children one might say that they were regressing over the summer. However, the point is that most of these relative effects are probably due to statistical regression. The top and bottom groups are selected from the extremes of the distribution at one time, and they are likely to contain some number of children who are there erroneously, i.e. whose true scores would lie in the middle of the distribution or at the other end. In the spring the same children are tested as a group and there will be some tendency for those included by error to have scores closer to their true scores the second time thus pulling the average of the lower group up, and the higher group down (and increasing the dispersion of all three groups). This would happen to any attempt to provide change norms for extreme groups when the pre-test scores were also used to select the extreme group. Obviously the point would be to select or sample the extreme groups (or any groups for which change norms are desired) on the basis of a different test - or on the basis of other group characteristics about which one wants to generalize, or for whom one wants to contrast the effectiveness of schools or programs, and then administer pre and post tests for these children (in an appropriate sample of schools or programs) for the periods for which one wants norms.

Put this way it is obvious why one should want to turn to control group designs to study change and relative effectivenss rather than trying to make do with norms. Still, if one insists on trying to use the norm tables to derive some expectations about the growth for a particular group, we would suggest using the change in the average scale score which would maintain the group at the same percentile from one point to the next. (In the MAT tables that figure varies considerably across percentiles, and it fluctuates from year to year for any given percentile - the differences in standard scores for Total Reading from 1st grade to 2nd, 2nd to 3rd, and 3rd to 4th for students at the 16th percentile are 16, 5, and 7 respectively.) We say that because we think the point of referring to norms is to ask how would the same kinds of kids do in general in schools in this country. Barring actually measuring the same kids more than once, the closest we can come to the same kids in the spirit of these tests and the assumptions on which they are designed are the kids who are at the same place in the distribution relative to their peers. We suggest measuring the amount of change in scale scores since, if their equal interval properties are accepted, then it might be argued that it would be appropriate to contrast the amounts of gain expected for different groups, or even to compare groups which start from different points in terms of the ratio of their actual gain to the gain which would be expected or normal for such a group (if there is equality of intervals

then it is appropriate to compare ratios of intervals).

But these are issues that also apply to control group designs. The point of using a control group is that if one can assume that the groups are similar except for the treatment, then the observed differences in outcomes or rates of change can be attributed to the treatment and the assumed similarity of the groups makes the uncertainty about the expected values, as revealed by the norms, unnecessary. The control group defines the expected value. Of course the appropriateness of the expected value so defined depends on the "typicalness," or other relevance, of the treatment given the controls (if the question is whether this school, or program, is better than average - the control treatment should represent the average), as well as on the initial similarity of the groups, and "typicalness" is probably even harder to measure than achievement. Still, those are design questions rather than metric ones, for our purposes.

The FTPV design involves the use of non-Follow Through classes which are at least similar to the Follow Through classes for the various sponsors. It has been argued for comparable designs that the availability of a control makes it unnecessary to be very concerned about which metric is used. However, the Follow Through and non-Follow Through groups were not picked by random assignment, nor were they systematically matched. Because of this they inevitably have differed in their characteristics and initial status,

and the national evaluation has had to use various devices to take these differences into account in comparing the effects of the Follow Through program to its non-Follow Through "controls," as well as comparing the effects of the sponsors to each other, where that has been done. These devices, analyses of covariance and various regression techniques, all involve assumptions about equal interval properties of the scales used, and we would make the same point here that we made above: For these purposes we see every reason to prefer the use of scale scores as the metric in such analyses. At least using these scores when the size of differences between pairs of groups or the relative amounts of change across groups are being compared, there is a basis for arguing that differences which are numerically the same are in fact comparable regardless of the starting point of the groups. With GE's, percentiles, and raw scores it is clear that the same numerical differences mean different things depending upon the starting point. In the middle range of scores the non-linearity between the various metrics may be relatively small, but as you move out in the distribution (and FT groups were presumably chosen because they were relatively disadvantaged) the possibility of distortions gets to be more serious.

There is another kind of problem which applies to the groups which fall toward the ends of the distribution. It is that it is quite likely that the measurement provided by these tests is

less reliable in those regions. The reason is that, as we noted, in making the tests the developers picked items which tended to bunch in the middle range of difficulty even though at least some items were included from the full range. If standard scores, scale scores, or percentiles are plotted on the X axis against raw scores on the Y axis the curve has an S shape; that is, small changes in raw score produce large changes in the other score at the ends of the distributions and relatively smaller changes in the middle. Thus, the precision of the discriminations made in mid-distribution is greater than at the ends, and small changes in the fundamental data - the raw scores - can produce relatively large changes in the scores for extreme groups. The scores are more subject to variation due to small errors, and hence they are less reliable at the extremes. We don't know how to estimate the potential impact of this effect on FTPV findings, and the publishers do not report varying standard errors for different points in the distribution, but the more extreme the groups being served by the programs being evaluated the greater one's concern should be about using these tests.

Finally, it has been pointed out that, given the non-linearity of the transformations employed in producing scale, GE and percentile scores (and hence the non-linearity of their relationships to each other), when the group means are being compared on GE's say, or percentiles, it is possible to get somewhat different numerical results depending on when one chooses to transform raw scores or

standard scores into the other metrics. One result can be obtained when the raw scores for individuals are transformed into the other metric and then the means are found in that metric, and a different result can be obtained if the means are found first in raw score terms and then they are transformed. Madaus (1974 p.23) reports that when he tried different orders of aggregation and transformation in a simulation he found differences of from one to four tenths of a GE in mean values, and in some cases a shift into or out of the zone of statistical significance for the observed differences. The size of the effect is, however, a necessary function of the degree of non-linearity between the two metrics and the extent to which the groups being compared come from the more non-linear sections of the distribution. That is, the distortion <u>can</u> happen, but it may not be a practical concern in some real situations.

To assess the practical significance of this concern we undertook an empirical check using FTPV data.

We used the data from the Philadelphia FT groups, Bank Street, EDC, and Kansas, and the Summer Study results in these analyses of whether the results would differ depending on the metric used, or whether varying the order of aggregation and transformation in the scores would make a difference.

The results in both cases are that these variations make essentially no difference in the significance of observed differences or the size of correlations.

Thus, while these concerns have a mathematical and necessary basis, their practical importance depends on the particular test, the relationships among the metrics for that test, and the region of the distribution in which one is working. With the MAT in the Philadelphia results there was no practical problem. With a different test or in a different area one should look carefully at the results and at the relationship among the metrics if it can be obtained from the publishers, or generated from their tables.

Nevertheless, as a general rule, again accepting the assumptions on which the tests are designed, we would recommend transforming individual raw scores to publishers standard scores and using those scores for all computations.

We don't see anything to be gained by generating standard scores based on the local distributions and using those scores for analyses. We recognize that that represents a kind of local norming and provides a metric for comparison of changes or differences between groups in terms of their joint standard deviation, but if the groups are extreme it does not escape the inherent unreliability of extreme scores, and it does not have as careful a psychometric rationale for the scale properties of these scores as the publishers have already provided for their scale scores. (These are also

some of the problems involved in exercises such as SDC's (1975) for the ESAA evaluation - which tries to develop "local" norms for the population of ESAA schools). Of course there may be good political reasons for reporting results in terms that can't readily be referred to national norms.

In general with control group designs there is an additional problem of interpreting the educational significance of the differences which may be found. Arbitrary solutions to this are often suggested - such as: an effective program is one which "moves" a group at least a third of a standard deviation relative to the joint distribution with the controls, or relative to the national distribution. But such suggestions <u>are</u> arbitrary, and they point out a fundamental problem with these tests - that it is not clear how what they measure is related to the things which citizens and policy makers care about when they assess schools. This is what we have called the "psychometric" problem with standardized achievement tests.

IV. PSYCHOMETRIC PROBLEMS INVOLVED IN THE USE OF STANDARDIZED ACHIEVEMENT TESTS FOR POLICY RESEARCH AND EVALUATION

The argument here is essentially about the validity of the tests - what they in fact measure.

What is wanted for these purposes is tests which measure the skills which schools <u>should</u> teach and which will show whether children have or have not learned these skills. Usually policy research and evaluation is concerned with a competition among

programs, and one wants to know which do better when measured against a common standard. The issue is differential effectiveness. But one can also ask whether programs accomplish different things, whether they have different goals and whether they achieve them. Here the issue is the measuring of differential effects.

Our first concern is with measuring differential effectiveness. We have shown that the claim for the validity of achievement tests as measures of the skills they say they are measuring rests on the quality of their sampling, with items, of the objectives of the curricula of U.S. schools and the content validity of the items - expert consensus that the items seem to measure what they are supposed to. The only additional outside tests of validity, once the tests are already designed, involve correlating the test scores with school grades (or with scores on other tests, if that represents an "outside" check).

But the basic validation of these tests is the correlation of the items with the total scores on the pool of items in the try out. The psychometric procedures select out of the pool of items those items which scale in the sense we described - i.e. they generate a distribution of scores, discriminate well among individuals, and "hang together." The fundamental criteria are distribution and correlation .

We argue that it is just at this point that the tests could

come to measure things other than the skills they purport to measure. It is the nature of any test item that it can require more than one skill to "pass" it. If the item is passed, one can argue that the child had to have the knowledge required by the item on its face (though even that is not necessary for some of the items in the tests - we can illustrate), but if it is failed, it is not possible to argue, for any one item, that the failure was due to the child's not having that skill. One of the other requirements of the item could also have caused it. That is, a math item may also require reading, a "reading" item may also require the use of inference; and the child who can do the computation required, or read the passage, may still not succeed on the item.

In general, performance on the items on the tests may require a great many other things - test taking skills, motivation to do well or to perservere, reasoning and guessing skills, broad incidental knowledge, and so on. If any or all of these things tend to cause children to differ from each other and if they cluster or scale, then they could be as, or more, responsible for generating the kinds of relationships among items which would lead them to be chosen by the psychometric criteria used by the test developers.

Such factors could be "general ability," or "white middle classness," etc., etc. They might be quite important, and they might

be worth teaching and measuring, but they are not the sorts of things that the achievement tests say they are measuring, and not the sorts of things schools usually are trying explicitly to teach.

It is hard to demonstrate empirically how serious this possibility might be, because the only real test would be to develop alternative ways of observing and assessing whether children actually have the particular skills the tests are supposed to measure and then checking to see how closely the test results match such alternative results. But this is precisely the work the test developers have not done.

A kind of check could be made by looking at item factor analyses. The sorts of "other" factors we have described ought to be relatively more common to all test items (of a given format) than would the specific skills. Thus if such factors predominate in children's performance on the tests the items ought to be as closely related across subject matter sub-tests in a battery as they are within a particular sub-test. It is interesting that the test developers seem not to have done such analyses themselves (though they do report the intercorrelations among sub-tests - and they are high). We checked with the publishers of the MAT and the CAT, however, and found that at least one graduate thesis has been based on such a factor analysis. The result (for the Stanford Achievement Test of HBJ) was that there was first a very large general factor and then,

to their relief, a series of specific factors which loaded for the most part on particular sub-tests. This seems to us to argue that our concern is real - the tests do "contain" or measure general ability, or some other general variable, to a considerable extent, but they also, to some extent, measure specific skills. Logically this makes sense, if one accepts the picture of the items as requiring multiple skills, including specific ones, for successful performance. In fact, for children who are high on the general skills, variability in the test should reflect variability in specific skills; for children who are low on the general skills, the variability could be more related to variation in those skills, and the tests might measure different things at different points in the distribution - or for different groups.

Our point is that we can never really know until a great deal more work is done on specific validation. And if such work is done, it would probably be better to use it to design new tests which are tailored to their specific purposes, as described in Section V.

The implications of this argument for evaluation and policy research are these: If these tests do measure general skills rather more than they do the specific skills which schools are trying to teach, and if children's differences on these general skills are more a matter of ability or background (absent explicit attempts

of schools to teach them), then the use of these tests in policy research should generate just the findings which Coleman and others have obtained - that variations in the outcomes as measured by these tests are more closely associated with variations in students' backgrounds than they are with measured differences among schools.

This implication could be tested if there were alternative tests available designed more carefully to be measures of specific skills. A Coleman style study run with such tests should show a greater proportion of the variance being "explained" by school variables when such tests are used than if the standardized tests are used (if, of course, the schools do truly vary in their effectiveness in teaching more specific skills). There are a few such tests available - "criterion-referenced" and "objective-referenced" tests, and the more traditional syllabus oriented tests like the old N.Y. State Regent's exam and the kinds of exams used in the U.K. There is research now under way to test these possibilities (Madaus and Airasian, Boston College; Kiesling and O'Reilly, New York State Education Department). Preliminary research from one such study (Madaus - Airasian, personal communication), indicate that the predictions may be borne out. That is, in Ireland using the Irish School Leaving Certificate (a syllabus related test) and a standardized test, the proportion of the variance which is explained by school factors is significantly greater with the School Cert. subject matter tests than with the standardized tests. These of

course are preliminary results, with a number of methodological difficulties still to be solved, but given the strength of the logical argument based on the nature of the tests' design and validation, we think they are sufficient to cause concern.

We doubt that further use of standardized tests as a basis for evaluating differential program or school effectiveness is justified at this time, if the concern is with the schools' effectiveness in teaching identifiable, specific skills. If the concern is with their effectiveness in changing general ability, one might as well use I.Q. tests, but we are not sure why that concern would be appropriate or fruitful at this point. It seems to us that it is time to be a great deal more careful about the specific, and general, goals of schooling and to find more defensible ways of measuring them, before a great deal more money is spent on evaluation research.*

---

*We should earlier also have mentioned that there are some obvious problems in the ways in which GE's and Standard Scores measure across school years. The point is that they can't measure the same thing, in terms of specific skills, in both directions. That is, a sixth grader who scores at the second grade level in a math test is not responding to the same kind of items as the median second grader was--visa versa for the second grader who scores at the sixth grade level. Their meaning has to be in terms of some more general concept of relative difficulty as measured by relative position in the group. This adds to the case that whatever the standardized tests measure it is something rather general. Under the scaling assumptions used it should be just about as easy to scale an eighth grade reading test against a second grade math test, as it would be to put two reading tests on the same scale. The psychometric fit would be likely to be just as good. If that is possible, it seems likely that the main things being "measured" by the two tests must be very similar.

The argument on the use of these tests to measure the differential effects of schools or programs is fairly simple. It is that in the process of designing the tests to be fair to all schools the test makers have written item specifications and chosen items which are intentionally made to be insensitive to the particular or idiosyncratic differences among schools, curricula, etc. This pushes the item content toward the more general, and makes them unlikely to pick up either differential effects, or differences in specific effectiveness.

Ideally to test the significance of this problem one would want to use the ability to discriminate among groups known to differ in various ways as a criterion for item selection in the design of tests which would identify such differences. Lewy (1973) has proposed something like such an approach and suggested the use of the intra-class correlation as the statistic, comparable to discrimination and difficulty indices, to use as a criterion. The intra-class correlation is an index, essentially, of the ratio of the variance between two groups "identified" by the item to the variance within the groups. It is mathematically closely related to eta, a product of analysis of variance.

As a first approximation of this procedure we undertook an analysis of the relative ability of the MAT items to discriminate among the Philadelphia Follow Through groups. The sponsors were EDC, Bank Street, and Kansas, and we assumed that the contrast in approaches between EDC and Bank Street on the one hand and Kansas on the other might generate some differences which would lead to differential discrimination among the groups by different MAT items.

Our results (See Appendix C) indicate that we were able to find MAT items which showed higher etas than did the test or subtests as a whole, and some of them differentiated the groups in directions other than the order of differences shown by the test as a whole. We attempted to make up new "tests" using items which discriminated among the groups in the same way (this represents a refinement on earlier attempts (Rakow 1974, Airasian, Madaus, and Rakow 1974) which tended to ignore the direction in which the items differentiated the groups - simply lumping the items which maximized discrimination into one test; we tried a number of ways of getting "purer" tests - picking the items which all discriminated in the same way, correcting the item difficulties for the overall performance of the groups and then looking at the direction of the discrimination, and factor analyzing the items and using only items which loaded on the same factors), and we did find some such tests in a given year which discriminated better than the test as a whole. However, the differences were not large, and when we looked at the results

for a second year using the same sponsors we found that none of the tests so devised replicated. We therefore think that this kind of secondary item analysis using items on standardized tests which have already been through the psychometric wringer is not likely to be fruitful (though one test, with groups whose "true" differences were uncertain, is hardly definitive).

In any case this sort of exercise seems to us to miss the main point of policy research and evaluation. The idea is not to identify any and all differences which might exist between schools or programs, but rather to identify those differences with which one is concerned - differences in effectiveness in instruction on important common objectives, or differences in effects which are intended by the program designers, when those differences in fact exist. That is, one needs valid tests designed to assess the things one is concerned about. The design of such tests requires a rationale which goes well beyond the kind of empirical search which has so far been suggested (and which the current tests represent).

Our argument has focused heavily on the point that the empirical methods used in test construction produce tests which tend to distribute students in a particular way, but what if we abandon empiricism for a moment and think about what ought to happen, or what we would like to have happen? If our concern is for the use of achievement scores in program evaluation or school accountability,

what should be the distribution of scores produced by schools or programs? It seems to us that the case can be made strongly that it ought not to be a normal distribution. Instead many teachers presumably should be trying to teach much of their curriculum to everyone in the class, and the result could be no distribution. Everyone would pass. Or they would expect everyone to achieve at least a basic competence in the subject. In that case the distribution of scores on some real scale of performance might be expected to be truncated at the lower end, even though it might still tail out at the higher end. A school one of whose goals was to reduce the inequality in the acquisition of school skills might also be seeking such a distribution (Klitgaard and Hall 1973).

These tests simply may not be able to yield such results. One of the complaints frequently heard about norm-referenced tests is that half of the population has to be below the median, and thus they are not good tests to use for evaluation and accountability purposes - it would be foolish to set median performance as a goal for everyone to reach. That of course is true if the reference were to the median of the population currently being tested. But if the reference is to the median score of the norming population established when the test was designed, it is not impossible for all the members of a subsequent population to exceed that

score. As long as the norm is not allowed to increase with increasing performance of new groups, it is not impossible. Norm-referencing per se is not the issue we are stressing here. The issue is rather that the distribution of scores on the test may be relatively impervious to being changed by school interventions which could in fact change the performance of children on the specific skills the tests are supposed to measure but which have little impact on the other skills which are actually producing the distribution. The other skills may also be important, and the tests might be useful for evaluation of programs which were aimed at changing them, but we are arguing that if the tests resist showing changes in the distribution of scores on specific skills when such changes "really" could have happened, they are not appropriate tests for use in evaluating instructional programs aimed mainly at changing achievement in those specific skills. As long as there is uncertainty about these questions, the tests will give ambiguous information when used for policy purposes. They won't allow us to determine which programs work or fail for encouraging particular skills, and they will provide ambiguous or bad information even for diagnosing the specific weaknesses or strengths of individuals.

V. RECOMMENDATIONS FOR THE DESIGN OF TESTS TO FIT THE PARTICULAR PURPOSES OF TESTING

Our argument up to this point has been designed to raise questions about the psychometric properties of achievement tests

and to justify as carefully as possible that such concerns could be real and could affect the kinds of results that will be obtained when such tests are used in policy research. Assuming that we have succeeded in justifying the concern, then the question becomes how to test whether it has real, practical effects and, if so, how we could develop alternative measures which would be more appropriate. We think the key to both parts of the question is the same. It is obviously a question of validity. What is required is that policy researchers and policy makers should be very clear about what effects and goals they wish to measure and what the purposes of that measurement will be. Then, if possible, independent criteria for telling whether those effects or goals have been obtained should be identified, and then the old tests can be related to those criteria to see whether they give the same results, or new items and tests can be developed which correlate with or predict to those criteria. If it is not possible to identify and get agreement on completely independent criteria, and if some reliance has still to be placed on empirical bootstrapping, then much greater care should be taken in making explicit the structure of assumptions and the logic of the relationships which allow one to say that the tests being developed measure what they are supposed to. This advice is cheap, and it requires some elaboration to make clear just how difficult its implementation may be.

The purposes for which policy makers and researchers might want to measure students' performance in the cognitive skills supposedly taught in schools include: assessment of what levels of skill students in or leaving school are attaining or whether they are attaining "adequate" levels of skill, in order to find out whether the schools should be encouraged to do better; evaluation, on some common basis, of the levels of performance or the adequacy of performance attained by students in different programs to see which programs ought to be adopted or extended; investigation of what the differences are in the effects of different programs; identification or diagnosis of particular stages, problems, or weaknesses in the development of students' skills and relating those diagnoses to particular instructional prescriptions; measuring the impact of instruction or, perhaps as part of the process of fitting prescription to diagnosis, assessing what changes occur over varying periods and with varying types of instruction; and, possibly, measuring levels of need for instruction as a basis for determining the amounts and kinds of instructional resources and time (and the allocation of funds) that will be required to reach particular instructional objectives.

Assessment and Evaluation

The first two purposes, assessment and evaluation in terms of common objectives, are clearly related. The first wants to

know whether we should view with pride or alarm. The second, if some level of alarm is sounded, serves the search for causes and cures using the same standards. In both cases the measures used should yield information about levels of performance which has some meaningful interpretation for policy makers and the public, particularly in distinguishing between levels of performance which are satisfactory and those which are not.

As we have tried to show, norm-referenced testing yields one kind of interpretable measure, but one which is interpretable only in terms of the distribution of performance of the reference groups. There are at least two other general strategies for obtaining interpretable measures. The first involves reference to a theory, and empirical knowledge, about the development or structure of a particular skill, and the second involves reference to real life performances requiring the skill, usually performance in various adult roles, and asks what items or measures of the skill can be shown to be related to (predict) direct ratings of those performances - either as being adequate or inadequate or as representing various levels of competence or worth (see Appendix B). It will be evident that the two strategies are not completely distinguishable in principle.

The theoretical appraoch involves working from an understanding of the skill area being assessed. For instance for reading, if we had a good developmental theory of reading which specified what its component skills were and which ones were necessary before others

could be acquired, and so on, we would have the basis for designing a set of tests aimed at measuring those component skills. If the acquisition of those skills were in fact ordered, and acquiring later ones were dependent upon acquiring earlier ones, it might be possible empirically to identify threshold levels on measures of the earlier skills which would indicate readiness to go on to the later ones. Such tests would yield information of quite direct usefulness both for the assessment of whether students were reaching the appropriate thresholds at appropriate times and for the evaluation of programs to see which of them were more likely to bring students to these levels. Whether or not the various skill areas are structured in a way that makes it sensible to talk about thresholds and contingencies is in part an empirical question, (math may differ in this respect from reading, for instance) but it certainly seems likely to be a fruitful one to ask. We do have a kind of theory of development of school subjects now, of course, as it is embodied in the curriculum of our schools, and even that could provide a basis for an ordered series of tests if somewhat more care were given to seeing that the tests measured what they say they do, and if they were somewhat more carefully related to at least the major sub-divisions of the skill areas. The expanded standard scores of the current tests, if change from year to year were allowed to be a more powerful criterion for item selection and if the psychometric impact of distributional considerations were reduced, would represent a

pragmatic beginning on this process. Some of the commercial criterion referenced tests (also see Appendix B) which have begun to be available are tied in this way to curriculum and have some of the attributes one would want for these purposes, but they tend to be too specific, both in terms of the size of the skill domain measured and in terms of their usefulness across curricula, to be ideal for giving publicly interpretable results.

The second approach to assigning a meaning to test scores which makes them useful for assessment and evaluation involves relating test performance to the requirements of other performances, particularly those involved in various extra-school societal or adult roles. This approach seems more likely to be taken in the later years of schooling, while the theoretical or developmental approach seems more likely in the earlier years. The problem, of course, is finding a convincing and valid basis for making judgements of the criterion performance - for instance what is a good or adequate mechanic, lawyer, or citizen? But if these judgements can be made, then it might be possible to select items and tests on the basis of whether they discriminated between people who perform adequately and those who don't or whether they predicted the ordering of criterion performance. Note in this latter case that if the prediction is to scaling based explicitly or implicitly on some other reference distribution (as another test, some school grades, etc.) it is indistinguishable from norm-referencing,

so we are talking here about ratings made on some other basis. The tests can be developed to measure and order skill in a specified domain, and then the question is whether and at what levels the skill is required for the criterion performance, or the criterion performance can be used to define the tests and items which relate to it on a more empirical basis. The latter technique would be likely to select aptitude type items and tap other dimensions in addition to specific skills. Since we are concerned here with assessment and evaluation of skills which schools are supposed to teach, and not with selection into these adult roles, we think that trying to measure the skills directly and then setting interpretable thresholds or cutting points on the basis of relationship to external criteria is the more appropriate approach. Work of this sort has only begun to be carried out with respect to school skills. The army and others have done some work on the relationship between reading and literacy levels and performance in some specific military jobs. ETS has been surveying the actual reading behaviors adults engage in, and developing tests based on them, presumably as a preliminary to saying something about what is required for adequate performance levels in literacy areas using test items for which adequate performance levels can be defined (for instance, can a person address a letter well enough so that it would go through the mails, or write a check that a bank would cash). Work of this

sort may well receive a spur from court decisions which are requiring that tests used for selection in situations where there is a clear difference in the rates of selection of majority and minority persons have to be shown to be related to actual job performance.

One of the points of developing tests on the basis of their relationship to external criteria rather than simply using the external criterion ratings themselves, of course, is that it may be relatively difficult to make the latter judgements. There is an obvious advantage to having a test which is relatively simple and less time consuming to administer but which validly stands for or predicts to the criterion performance. The judgements can be made validly and expensively in the test development process, and the test can do the work from then on. With suitable safeguards and vigilance against self-fulfilling prophecies, this is a reasonable argument. It raises one very serious caveat, however. We suggested earlier that one of the reasons that standardized test items may not be valid measures of what they purport to measure is that they contain requirements other than the particular skill supposedly being tapped. Some of these requirements are probably directly related to the item formats and the conditions of test taking, which are themselves conditioned by those practical considerations of time, economy, convenience, machine scoreability, etc. which lead us to use tests rather than direct observations of the criterion behaviors. This problem is compounded when the

overall test scores become a key criterion for item selection, as they do in many standardized tests, so that there is no independent check on validity at all. But it will not be easy to escape even when there is an attempt to develop independent criteria. What is required in the case of academic achievement tests is a period of extensive experimentation with alternative ways of eliciting the desired skills - to see whether children who fail to show the skill in the multiple choice test setting, with its dependence on reading and on "psyching-out" the test maker, will show it in other settings - in games, in conversation, in more casual or less time pressured situations, etc. (Lazarus 1973, from unpublished proposal). Then it can be seen whether machine scoreable items can be written which can show these skills when they are known to be present from other evidence. Perhaps tests will have to be devised which give children alternative ways of showing they have a skill (Lazarus 1973, unpublished proposal), and judgments will have to be made, if children show skills in some situations and not in others, about which situation-skill contingencies are the ones which schools should be held accountable for producing and children will have to be able to do to get along in this society.

Differentiating Programs

In some cases policy makers will want to know not so much whether programs differ in their effectiveness measured against

some common standard but rather how they differ in general - what are the differences in the results they produce and are those differences ones which the programs said they would produce? If programs do in fact have different effects, and if policy makers, communitities, parents, etc. differ in their preferences for outcomes in ways related to those differences, then this information could be important to them in maximizing the fit between the programs and their preferences, and in telling them what they were trading off in making those choices.

Earlier we argued that standardized tests are likely not to reflect particular program differences, partly because they avoid items that are specific to particular programs and partly because they may not really be measuring the skills which are the focus of school programs but instead may be tapping general aptitudes, or class and cultural differences, etc. which are less likely to be affected by schools. If we had tests of school skills which most people would agree measured those skills then the fact that they were free of the idiosyncracies of particular curricula would be an advantage, especially when used for accountability. It is because people, including school people, may have differing ideas of what achievement in particular skills means or because they have differing notions of what the schools should teach that we might want to devise tests which either reflect or define those differences.

Again there are at least a couple of ways to go about this. First if the programs themselves are well defined, in terms of theory or their pedagogical approach, it may be that they can specify how they should differ from others in sufficient detail so that tests could be developed to measure those goals. Any or all of the techniques for test development mentioned above could be used, and the tests so developed could be used both to test whether the programs were doing what they hoped they would and to see whether other programs, with or without the same goals, might be doing just as well. Obviously such a strategy could go well beyond strictly cognitive goals, and it is one that might have been followed in FTPV if the sponsors had had the time, resources, and clarity of purpose to pursue it.

The second approach is more simply empirical. It asks, given a set of programs which presumably differ, what items or tests differentiate among them? In effect it takes whatever differences exist among the effects of the programs and uses them as the criteria against which to validate items or tests. If items or tests can be found which do systemmatically and consistently differentiate among the groups, they might then be inspected to see whether they have any clear interpretation or significance. If they do, this process would begin to define what the programs were in fact about, and the meaning of their differences. Presumably this strategy

would make the most sense if it were pursued as broadly as possible, with a wide range of types of items and tests, rather than with the predigested items available on current standardized tests, but the range could be narrowed by having good guesses about what the nature of the differences is likely to be. Even a new look at the old test items in this light might be of some interest, if only to test the suspicion that they don't differentiate very well, as our own work has implied.

Diagnosis, Prescription, and the Effects of Instruction

The business of designing and validating tests which are useful for telling where a student is in the process of developing a particular skill, identifying the kinds of problems he may be having, and specifying what instructional approach would be most likely to move him or her forward involves a great deal of work and a long process of trial and error. It would be facilitated if there were decent theories of development and structure in these fields of the type described above, which could guide both the development of items and the selection of instructional approaches. There are now only the beginnings of such theories, if that. There are, as we noted, the theories implicit in the current curricula and in the lore and experience of teachers, and that basically is where the process will have to start. It involves: deciding what the likely stages and problems are in the development of the skill; writing

items which seem to tap them and trying them out; seeking whatever form of validation is available; trying various instructional approaches to changing the students' performances; and seeing whether the tests and items, or which of them, show changes in conjunction with instruction. If the tests fail to show changes, either they should be abandoned, or the instructional techniques should be reconsidered, or both. Which conclusion is drawn will depend on the degree of faith, or other validation, that has been built up for the test or the form of instruction, and the process will have to iterate again and again before it generates much certainty. If some items reflect change with instruction and others don't, or if they change with some instructional techniques and not with others (or with differing periods of instruction or with different children), inspection of the items and techniques to see how they differ could begin to build up an empirical basis for understanding the structure of the skill domain and the nature of learning within it. What we are talking about of course is the fundamental process of education, with a little help from our tests, and there is no reason why answers should come easily, quickly, or permanently.

Educational Need

The final area we wish to discuss here, in which policy makers may find tests useful, has to do with measuring students' educational

needs in a way that would indicate the kinds and amounts of educational resources required to produce particular results for them. The ultimate answers to this question will await the results of the kinds of investigation we described with respect to diagnosis, prescription, and instruction, since that should if successful yield an understanding of what instructional techniques work with which students, and presumably the costs of those techniques could be specified and a predicted cost of instruction associated with each student or type of student. However, it may be that this is one area in which norm-referenced measures have a clear usefulness, since if it were true that they tapped either relative ability to learn or the relative degree of cultural difference between the school and the child, either one of these might yield at least a crude index of the relative special effort schools might have to make with students. This could provide a basis for a tentative extra allocation of resources or funds, which could stand until better information became available from the longer process described above.

It would be better, however, if measures of ability to learn or likely speed or ease of learning are desired, to try to develop tests validated against speed or ease of learning as measured in actual learning situations. Current aptitude or achievement tests might prove to be valid by such a criterion but the work necessary

for a direct validity check (as opposed to the more inferential one of relationship to school grades) has never really been done.

Appendix A

The Army Alpha

The discussion in the text of the Army Alpha and its development is rather idealized, and perhaps anachronistic. While most of the principles of test design described here were thought of by the psychologists who designed the Alpha, they were not applied as thoroughly and formally in the development and use of the Alpha as this discussion may imply. A contemporary and fascinating history of the Alpha, and related tests, is available, written by some of the participants (Yerkes, 1921). The tests were put together and tried out in an amazingly short time. Many of the ideas about group testing, standardization, objective answers simple responses and scoring, and a number of specific sub-tests were adapted from the recent work of Arthur Otis. From other writings (Otis, 1917, 1918) it is clear that Otis had thought out most of the basic issues of item analysis (appropriate range of difficulty, discriminating power - which he called "diagnostic value," etc.), and he had the basic concepts involved in equipercentile equating methods and the use of the assumption of the normality of the underlying distribution as a basis for arguing that normalizing transformations of scores would produce equal interval scales. In the haste with which the Alpha was put together, however, there wasn't enough time for a full item analysis for each of the eight sub-tests in the Alpha. Instead the sub-tests themselves were subjected to analyses, rather like item analyses, concerned with their

relative difficulty and relationships to each other and to the overall scores on the field test (the "statistical team" at Columbia University who were sent the results of the two week or so summer try-out of the first forms of Alpha was headed by E.L. Thorndike and included L.L. Thurstone!). There was some attention given to whether the different sub-tests' scores should be weighted differently in the total score (it turned out that different weights didn't seem to make much difference in the correlation of the total scores with other indices), and some sub-tests were kept because they seemed to be tapping different components of intelligence, while one or two were dropped, even though they behaved well statistically, because they seemed to be giving redundant information. The modifications made after the first try-out were slight, however, and the group moved to wider-scale field testing. In this they did pay a good deal of attention to correlations of the test scores with other kinds of performance ratings (and with Stanford-Binet I.Q. test scores - Terman found a range of correlations from .58 to .88 in a number of samples, with most of them in the .80's (Yerkes, ed. 1921, p. 332)), and they checked to see that the mean scores obtained when the tests were given to different military groups (officers, recruits, labor batallions, etc.) showed the appropriate rank orders. As mentioned below, these relationships were <u>not</u> used as criteria for item selection, or for significant test revision. The relationships to

other tests were taken more seriously as certifying that they were measuring intelligence, than were the relationships to military levels and performance. It is clear that the test makers were pretty confident that the tests were good measures of an important aspect of human ability and that if the measures didn't relate perfectly to real world functioning, that was probably because real world functioning required other skills as well (Thorndike is quoted (Yerkes, ed. 1921, p. 317) as noting that the ability measured by the tests, which was as much inborn as it was determined by circumstances, was "obviously specialized for words, numbers, and other abstract and symbolic content. Hence, the need of tests of ability with things and people before deciding the fate of a recruit.").

There was some attempt made to scale the scores for the Alpha, particularly in equating and weighting the sub-tests for difficulty, but the scores actually used by the military seem to have been reported mainly in terms of letter grades (A, B, C . . etc.) and, to some extent, in terms of mental age equivalents. The test manuals provided tables for translating score ranges into these terms, which were somehow more directly meaningful to the military. A more elaborate scaling, of the normalizing, equal interval, type was done for a companion individualized test, which was supposed to be used for doubtful cases identified by the

Alpha and Beta (i.e. for more definitive "detection of low-grade intelligence" and for more "accurate grading of foreigners," (Yerkes ed. 1921, p. 310). The test does not seem to have been widely used during the war, though, like Otis' test (Otis, 1918), in its psychometric procedures it may have been a closer ancestor of the national group intelligence and achievement tests developed right after the war.

The tests were introduced on a wide scale in the military in 1918, a great many men were tested, and undoubtedly the tests played a considerable role in military selection. The war ended fairly soon, however, and there was less chance than would have been ideal to follow up the selection decisions and to add the kind of validating information such follow up would have provided. Nevertheless the tests were widely acclaimed as a success, and this example of applied psycho-social "engineering" gave overwhelming momentum to the testing movement.

We have stressed here the consequential nature of the Army selection decisions as a basis for the motivation to design tests which could distinguish reliably among individuals, and that undoubtedly did play a significant role. The tradition of concern for individual differences was deeply rooted in the testing movement, and in its hereditarian origins, well before the war, and it would probably be more accurate to say simply that these psychologists, like their predecessors, thought that individual differences were themselves <u>important</u> and that tests designed to identify them should do so as <u>precisely</u> and reliably as possible.

Appendix B

Notes on the Norm-Referenced, Criterion-Referenced Distinction

It has become common to contrast norm-referenced testing with "criterion-referenced" testing. We think that the basic contrast with norm-referencing has to do with the origin and interpretation of the scores given to performances on a test. Norm-referenced scores - the numbers assigned to the performances - bear some pretty direct relationship to the distribution of performances in a reference population, and their basic meaning derives from that relationship. The general contrast would be to find some (any) other basis for assigning numbers to the performances on the test and for attributing meaning to those numbers (or to the test performances themselves).

One such basis involes somehow making a direct judgment about the level of test performance which can be considered adequate or acceptable, without reference to how many people can achieve that level. The judgment could be made in terms of the test performances themselves, or in terms of some other behavior against which the test performance is to be validated. In a somewhat more complicated case one might also be interested in making judgments about additional levels of performance beyond adequacy (in either direction). Whatever the number of levels to be identified and judged, a person's score would be related to whether or not it reaches any of those levels or is correlated with behavior which reaches any of those levels. If it does, it meets the standard or "criterion" for

adequacy, acceptability, or goodness at that level and has its meaning in terms of whatever basis was used in making the judgements about the levels. This is the essential meaning of "criterion-referenced" as applied to tests. If norm-referenced tests are seen as having meaning in terms of what you expect (in the statistical sense of "expected value"), and criterion-referencing as dealing with what you want, perhaps the psychometric contrast is the most fundamental one. Actually, the term can be applied even when the judgments are influenced by "reasonable" expectations about what people can do. The point is that they are freed from having to use the average, or any particular aspect of the distribution, as the standard.

It should be noted that criterion referencing in this sense "assigns numbers" to test performance in a quite limited way. Unless the test performances have also been scaled in more elaborate ways, of which there is brief mention below, the scores given to criterion-referenced tests are equivalent to giving zeros and ones, representing either failing or passing a simple criterion, or, if judgments are made about a number of levels of performance, the numbers associated with those criterion levels would at best have the properties of an ordinal scale. That seeming loss in scale sophistication in comparison with scores based on norm-referencing may not be a real loss in tests used for policy purposes, if the focus on the issue of setting appropriate criterion levels forces

attention to the validity of the tests for the particular uses to which they will be put. As we have argued in the report, the usual norm-referencing procedures allow the real issues of validity to be slighted or ignored.

Examples of criterion referencing tend to run to single level threshold judgements, often in physical skill areas where passing or failing is relatively unambiguous - as for instance sinking or swimming, riding a bike or falling off, etc. Where cognitive skills are concerned, examples tend to be rather specific - at the item level, or sets of very similar items, where passage or failure again is pretty unambiguous. In fact, much of the discussion of criterion-referenced tests treats them as though it is a necessary characteristic that they deal with relatively specific, closely curriculum-related, skills. If our definition is accepted, however, specificity isn't a necessary characteristic but rather a matter of convenience (or laziness). For instance, there is a good deal of talk now about scales of basic or functional literacy. Such scales could deal with rather broad sets of skills. As long as some joint function of the performances on the scales is defined as constituting an acceptable level of literacy, they could be considered criterion-referenced. It is the final summary judgement of the performance which has to be discrete, and in that sense specific, not the performance itself. How the judgements of acceptable levels would be made for complex skills such as literacy is of course a longer story.

If one sticks to the item level he is on clearer, if less useful, ground, but he is also in danger of missing the psychometric point. After all, the items on a norm-referenced test are also passed or failed. The issue is how you tell that passing or failing an item, or a set of items, represents something significant about a skill that you care about. With limited exceptions you can't tell just by looking. Much of the discussion of criterion referencing, including that done in connection with the National Assessment, does seem to miss this point.

Some writers on criterion-referenced tests seem also to think that because such tests do not derive their meaning from reference to a statistical norm they somehow also escape normative judgements and invidious comparisons. But, clearly, making a judgement about whether a performance is adequate or not is the essence of normativeness. And if you can make comparisons between groups on the proportions of their members who pass, you can hardly avoid comparisons which will be invidious in someone's eye. Moving away from norm-referencing allows you to set a criterion level which more than half of a group can pass, if you have some reason to do so, and it means that the test defines the group rather than the other way around, but definition is not escaped. Probably the desire to escape comparisons is also part of the reason for the emphasis on curriculum- and program-specificity in criterion referencing. If the test can only be used

to measure the results of your program, it can't tell whether you are doing better or worse than some other program. At least in principle, however, common goals can be set which different programs may or may not meet, and common criterion levels can be set to see whether they do.

"Objective-referencing" is a relative of criterion-referencing in which the domain of performance in which students are supposed to achieve is indicated at some degree of specificity, and items are written which seem to tap that domain. Sometimes a criterion level is set, at least in terms of the proportion of items which have to be passed to meet it, but sometimes the use of the term "objective" is a clue that explicit criterion levels will not be set and that the enterprise is more vaguely concerned with assessing how students do in an area without saying ahead of time what results would make us happy or unhappy.

Other bases for ordering and interpreting test scores exist, though it is easier to talk about how they order than it is to distinguish fundamentally different ways of assigning meaning. Sometimes these methods are said to yield "absolute" measures as opposed to the relative measures given by norm-referencing, but absolute is a slippery word here, and rather than get into an extended discussion of levels of scaling, we will stick to talking simply about bases for scaling other than reference to the score distribution. These

bases for ordering tests scores usually involve some consideration of item difficulty or of the contingent relationships among items (note, even here, that items tend to be of the form where some judgement - whether they are passed or failed - must be made about them, so that criterion referencing is built in at the beginning at the lowest level, though there are of course kinds of items which don't have right or wrong answers). Raw scores are a simple form of this. A more complicated approach would be to order items by difficulty and score in terms of the most difficult item passed. This tends to give ambiguous results unless a further complication is added in the manner of Guttman scaling, in which one looks for sets of items where the relationship of difficulties is well-ordered so that if one knows the item of greatest difficulty a student has passed he can be sure that all easier ones are passed. Even more elaborate patterns of contingencies among items or scales can be considered as a way both of learning about and of measuring the structure of a domain, and these relationships can be compared with the structure of more directly observed behaviors (Airasian, 1973). It is this latter kind of work - the investigation of structure and relationships - which offers an additional basis for meaning in that it is fundamental to generating a theory of behavior in any area. Such a theory may provide a kind of meaning for scores which is different from normative judgements based either on expectation or desire, but there are

not many good examples of this at present and it is hard to give a feeling for it in the abstract. Now, in most cases, the ordering provided by the tests will have to be related ultimately either to criterion judgements or to distributional norms.

If you think very far about these issues you are forced to recognize that in matters of meaning you come back ultimately to human judgements. The empirical techniques can identify and make visible new dimensions. The normative and psychological meanings of positions on those dimensions have to come from people. This observation suggests that we might do well to turn to the controlled use of scaling techniques which rely directly and explicitly on human judgements - such as the techniques of magnitude estimation worked out by Stevens for scaling sensory judgements. These techniques might be applied most appropriately to the criterion behaviors which tests are supposed to predict to, or be more convenient proxies for. Tests could then be given to the same population and validated against those criteria.

Appendix C

Results of the Item Analyses on
the Philadelphia Data

TABLE 1. List of items from Primary I and Primary II, by year. Shows only those items where the F-test (item by sponsor) is significant at .01 and in addition where the item orders the groups "213" ie. Kansas over Bank Street over EDC. For each item the values of eta are shown (in this case the ratio of between sums of squares to the total sums of squares) and the order of eta within this group of items ie. under the column order "1" indicates the highest eta, "2" the second highest eta, etc. Tied etas are treated arbitrarily.

PRIMARY I

Subtest:	Spring 1972			Spring 1973		
	Item	eta	order	Item	eta	order
Word Knowledge						
	02	031	39	→ 02	069	8
	03	060	19	08	062	12
	14	087	6	11	075	6
	20	036	36			
	26	038	40			
	30	037	34			
	31	077	9			
	34	073	10			
	35	039	31			
Word Analysis						
	01	047	24	05	062	13
	02	120	3	06	054	16
	05	043	27	09	054	15
	09	032	38	10	048	19
	15	058	21	12	064	10
	16	175	1	13	045	21
	18	105	5	15	067	9
	19	070	13	16	111	2
	20	077	8	19	039	30
	21	041	30	21	090	4
	24	072	11	23	090	5
	26	031	33	25	046	20
	32	070	14	26	051	18
	35	068	15	29	074	7
	36	061	18	30	041	27
	37	084	7	32	123	1
	39	054	23	35	053	17
Reading						
	07	034	37	16	044	25
	08	056	22	20	041	28
	28	045	26	33	103	3
	32	036	35	41	042	26
	38	106	4			

TABLE 1. continued

Math	Item	eta	order	Item	eta	order
	27	042	29	05	038	31
	31	038	32	14	041	29
	32	064	17	17	045	24
	33	031	41	27	036	32
	50	042	28	40	036	33
	54	072	12	45	045	23
	55	047	25	47	057	14
	56	059	20	55	064	11
	61	156	2	56	045	22
	62	065	16			

TABLE 1. continued. (NB: the order is given here for the first 20 items only)

PRIMARY II	Spring 1972			Spring 1973		
	Item	eta	order	Item	eta	order
Word Knowledge						
	07	062	7	06	054	
				07	119	2
				22	046	
				31	049	
Word Analysis						
	25	038	17	01	043	
	31	044	14	04	056	20
	34	048	11	05	040	
				11	033	
				15	070	15
				16	038	
				23	050	
				25	104	4
				28	038	
				34	042	
Reading						
	34	046	12	20	049	
				30	064	17
				33	085	9
				34	050	
Spelling						
	11	077	3	02	091	7
	13	045	13	04	058	19
	14	058	8	07	042	
	15	041	16	11	077	12
	17	066	6	14	074	14
	21	037	18	23	054	
	23	054	9	25	090	8
	26	037	19	29	075	13
	27	068	5			
	29	088	1			
Math computation						
	03	037	20	09	039	
	22	073	4	10	098	5
	30	077	2	13	050	
				15	097	6
				17	052	
				22	055	
				26	068	16

TABLE 1. continued

	Item	eta	order	Item	eta	order
Mth concepts						
	33	052	10	02	041	
	35	044	15	07	042	
				14	039	
				28	046	
				39	047	
Math problem solving						
				20	059	18
				21	114	3
				23	078	11
				24	047	
				25	082	10
				35	152	1

TABLE 2. This table is the same as Table 1 with one very important exception, the order of the groups is "123" here. Please note: the order is "213" in Table I. There are less than 20 items with statistical significance of .01, therefore, there are less than 20 items in the table.

PRIMARY I

	Item	eta	order	Item	eta	order
Word knowledge						
	04	019	6	15	036	4
	10	021	4			
Word Analysis						
	08	032	1			
Reading						
				17	029	6
				38	038	3
				40	046	2
Math						
	13	023	3	11	031	5
	14	020	5	12	046	1
	28	028	2			

TABLE 2. continued

PRIMARY II

	Item	eta	order	Item	eta	order
Word knowledge						
	06	037	16	11	041	12
	18	042	10	18	030	
	19	072	2	19	036	16
	22	031		24	039	14
	28	042	11			
	33	067	3			
	38	041	12			
Word analysis						
	02	038	15	03	099	
	08	051	5	12	130	1
	09	038	14	19	034	20
	10	027		20	034	19
	11	042	9	21	068	5
	16	063	4	30	042	11
	18	021				
	29	031	20			
Reading						
	14	031		17	027	
	15	043	6	37	042	10
	16	042	7			
	27	032	19			
Spelling						
	10	039	13	01	087	2
	18	027		03	033	
	22	032	18	05	027	
				08	060	6
				12	034	18
				17	080	3
				20	027	
				21	035	17
				24	038	15
				26	040	13
Math computation						
	12	024		03	043	9
Math concepts						
				03	043	8
				08	031	
				32	050	7
				34	069	4

TABLE 2. Continued

Math problem solving

12	103	1	17	030
17	033	17		
18	021			
28	024			
34	042	8		

TABLE 3. Analysis of subtests. Subtests were chosen from the total pool of items in the Primary I and Primary II tests, the procedure carried out separately for each year and each test, thus, four groups of subtests. The first subtest is one item: that item with the highest eta which also orders the groups "213" (see table 1.). The second subtest includes the item in subtest 1 and adds the item with the next highest eta. And so on, The first ten subtests differ from one another with respect to the last entered item only. The last subtest includes 20 items. Table shows eta for F-test (subtest by sponsor).

PRIMARY I	1972	1973
Subtest 1	175	123
2	297	192
3	369	253
4	321	259
5	361	249
6	353	253
7	311	271
8	290	257
9	290	242
10	286	241
.		
.		
20	262	206
PRIMARY II		
1	088	150
2	165	263
3	146	259
4	163	261
5	156	260
6	159	260
7	183	262
8	182	240
9	176	230
10	182	221
.		
.		
20	194	308

TABLE 4. Analysis of total tests. The tests here are made by summing all correct responses within the total pool of items (Primary I, Primary II). Values are eta for F-test, (Test by sponsor)

	1972	1973
PRIMARY I	118	107
PRIMARY II	052	086

TABLE 5. Analysis of subtests. Subtests were constructed in the same way as they were for Table 3 analyses EXCEPT that here the items all order the groups "123" rather than "213". Because there were fewer than 20 items for Primary I which had statistically significant eta (see Table 2), the subtests for Primary I do not get as big as they do for Primary II.

PRIMARY I

Subtest 1	032	046
2	053	074
3	063	072
4	070	087
5	070	088
6	076	114

PRIMARY II

1	103	130
2	138	169
3	103	146
4	124	195
5	147	216
6	162	196
7	154	216
8	161	209
9	146	188
10	155	185
.		
.		
20	103	111

TABLE 6. Subtest analysis. The selection of items for these subtests follows the procedures used in Table 3 analyses with one exception. An attempt is made to select only those items which are positively correlated with one another. Since all the items are positively correlated this means selecting items which are more inter-related than others -- ie. items which cluster. To find which items cluster all items for the Primary I subtest (1972) which appear in Table 1 were factor analysed. The factors were rotated using the varimax solution. The purpose here was to divide the variance explained by the factors into several distinguishable rotated factors, rather than let most all of the variance be accounted for by the principal component.

The first two rotated factors were used for the selection of items used in these analyses. Any item which had a reasonably large loading on one or other or these factors was dubbed a subtest item. Reasonable here means bigger than .3. Items were then added into subtests in order of their original eta. There are only 6 items in the second factor subtest.

PRIMARY I (1972)

	Factor 1 subtest	Factor 2 subtest
Subtest 1	105	156
2	123	140
3	147	144
4	187	127
5	162	119
6	175	137
7	169	
8	165	
9	156	
10	146	

TABLE 7. List of items from Primary I and Primary II by year. For these items the F-test (item by sponsor) is significant at .05 -- note: in Table 1 it is .01 -- and each item orders the groups "312" -- note: in Table 1 the order is "213". Where raw scores have been used in previous analyses, here deviation scores are used. A deviation score is the raw score (0 or 1) for the item minus the overall difficulty level for that test and that sponsor. Thus, deviation scores are calculated separately for each test and each sponsor and then merged to form a single variable.

The table shows eta from the F-test, and the order of the eta, 1 being the highest etc. Only 11 items were found with the required characteristics for each test.

PRIMARY I

Subtest	1972			1973		
	Item	eta	order	Item	eta	order
Word knowledge				09	036	7
Word analysis	04	030	7			
Reading	20	030	6	31	039	5
Math	02	097	2	01	042	4
	04	063	3	02	072	2
	12	143	1	03	089	1
	29	046	5	04	065	3
	30	051	4	10	024	11
	36	029	8	15	024	10
	37	027	10	36	031	9
	40	029	9	38	037	6
	41	022	11	62	034	8

TABLE 7. continued

PRIMARY II

Subtest	1972			1973		
	Item	eta	order	Item	eta	order
Word knowledge						
				33	068	5
Word analysis						
	17	040	6			
Reading						
	04	041	5	05	028	10
	05	061	3	09	113	2
	29	025	13	13	029	9
Spelling						
Math computation						
				01	080	4
				02	026	11
Math concepts						
	01	068	1	12	040	7
	02	033	8	13	094	3
	06	039	7	30	170	1
	09	057	4			
	11	025	12			
	16	027	9			
	30	063	2			
Math problem solving						
	07	025	11	01	042	6
	35	027	10	03	036	8

TABLE 8. Analysis of subtests. Subtests based on deviation scores for items which order groups "312" -- otherwise the description of Table 3 applies here too.

	1972	1973
PRIMARY I		
Subtest 1	143	089
2	206	128
3	224	130
4	192	129
5	203	142
6	211	141
7	206	133
8	196	151
9	169	143
10	152	137
11	136	138
PRIMARY II		
1	068	170
2	081	222
3	129	282
4	163	270
5	173	245
6	190	257
7	209	236
8	216	248
9	217	245
10	213	230
11	224	226

REFERENCES

Airasian, Peter W., Madaus, George F, and Rakow, Ernest A., Test Sensitivity in Program Evaluation, unpublished paper Boston College 1974.

Beck, Michael D., Research Psychologist, Harcourt, Brace, Jovanovich, Inc., Personal Communication on the "Fall-Spring Growth Study", July 15, 1974.

Bloom, B.S.,ed., Taxonomy of Educational Objectives Handbook I., The Cognitive Domain, New York: David McKay, 1956

Chronbach, Lee. Test Validation, Chapter 14 in Thorndike (ed.) (1970)

Coleman, James S., and Karweit, Nancy L., Measures of School Performance, Rand Corporation Document R-488-RC, Santa Monica, July 1970.

Flanagan, J.C., The Cooperative Achievement Tests: A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores. New York: American Council on Education, Cooperative Test Service, 1939.

Harcourt, Brace, Jovanovich, Inc., Metropolian Achievement Tests Special Report (1970 Ed.), Report No. 1, Content Development, HBJ, Inc.,New York, June 1971.

Harcourt, Brace, Jovanovich, Inc., Metropolitan Achievement Tests Special Report (1970 Ed.), Report No. 6, The National Item Analysis Program, HBJ, Inc., New York, June 1971.

Harcourt, Brace, Jovanovich, Inc., Metropolitan Achievement Tests Special Report (1970 Ed.), Report No. 8, Summary Statistics for National Standardization Groups, HBJ, Inc., New York, June 1971.

Harcourt, Brace, Jovanovich, Inc., Metropolitan Achievement Tests, Guidelines for Test Use and Interpretation, Guidelines No. 1, Development of the Standard Score System for the 1970 Edition of the Metropolitan Achievement Tests, HBJ, Inc., New York, October 1972.

Harcourt, Brace, Jovanovich, Inc., Metropolitan Achievement Tests Special Report (1970 Ed.), Report No. 21, Interpolated Midyear Norms for Grades 2 Through 7, HBJ Inc., New York, May 1972.

Harcourt, Brace, Jovanovich, Inc., Metropolitan Achievement Tests, Guidelines for Test Use and Interpretation, Guidelines No. 2, Development and Use of the Grade Equivalent Scale, HBJ, Inc., New York, February, 1973

Henryssen, Sten, Chapter 5 "Gathering, Analyzing, and Using Data on Test Items" (Thorndike, 1971).

Klitgaard, Robert E., and Hall, George, K., Achievement Scores and Educational Objectives, Rand Corp., Santa Monica, February 1973.

Lewy, Ariel, Discrimination Among Individuals v. Discrimination Among Groups, J.Ed. Meas., 1973, 10, 1, 19-24.

Loret, Peter G.; Seder, Alan; Bianchini, John C.; and Vale, Carol A., Equivalence and Norms, Tables for Selected Reading Achievement Tests (Grades 4,5,6) Anchor Test Study, 74-305 U.S. Department of Health, Education and Welfare, G.P.O.,

Madaus, George F., Memorandum to Members of the Joint Committee on Test Standards Revision - concerning the need for a comparison volume (on the uses of tests in evaluation), unpublished. Boston College, Newton, May 7, 1973.

Otis, Arthur S., A Criticism of the Yerkes Bridges Point Scale, with Alternative Suggestions, J. Ed. Psychol., March, 1917.

Otis, Arthur S., An Absolute Point Scale for the Group Measurement of Intelligence, J. Ed. Psychol., 9, May-June, 1918.

Rakow, Ernest A., Evaluation of Educational Program Differences via Achievement Test Item Difficulties, unpublished paper presented at AERA, Chicago 1974.

Stevens, S.S., Mathematics, Measurements, and Psychophysics in S.S. Stevens (Ed.) Handbook of Experimental Psychology, New York: Wiley, 1951, pp. 1-49.

System Development Corporation, Achievement Test Restandardization - Emergency School Aid Act National Evaluation, S.D.C., Santa Monica, 1975.

Thorndike, Robert L., ed., Educational Measurement, American Council on Education, Washington, D.C., 1971.

Thurstone, L.L., Primary Mental Abilities, Psychometric Monographs, 1938, No. 1.

Wisman, Alexander G., Chapter 4, "Writing the Test Item" (Thorndike, 1971).

Wrightstone, J.W., Hogan, T.P., and Abbott, M.M., Accountability in Education and Associated Measurement Problems, Test Service Notebook 33, Harcourt, Brace, Jovanovich, Inc., New York.N.D.

Yerkes, Robert M., ed. Psychological Examining in the United States Army, Memoirs of the National Academy of Sciences, Vol. 15, Washington, G.P.O., 1921.

Yoakum, Clarence S., and Yerkes, Robert M., Army Mental Tests, New York: Henry Holt and Co., 1920.